Metamemory across the lifespan:
Relations between memory self-reports and memory performance in older adults

Thesis
presented to the Faculty of Arts
of the
University of Zurich
for the degree of Doctor of Philosophy

by
Philippe Rast
of Hochdorf (LU) and Zürich

Accepted in the autumn semester 2007 on the recommendation of
Prof. Dr. Mike Martin and Prof. Dr. Friedrich Wilkening

2007

Acknowledgment

First of all, I would like to express my gratitude to Mike Martin and Daniel Zimprich for their mentoring, for all scientific support in every phase of this thesis and beyond. Further, I would like to thank Mike Martin and Friedrich Wilkening for being the referees of the present thesis and Mathias Allemand and Myriam Dellenbach for the support they gave my during the whole phase of the dissertation. Special thanks go to Elena and Quirin who contributed essentially to the success of this work by their own ways.

1 Introduction 1
1.1 Aging and memory performance 1
1.2 Aging and metamemory 2
1.2.1 Memory self-reports 4
1.2.1.1 Measurement of memory self-reports 6
1.2.2 Age-related changes in memory self-reports 7
1.3 Relation between memory self-reports and memory performance 9
1.3.1 Explanations for the low relation between memory self-reports and objective memory performance 11
1.4 Open research questions on memory self-reports and memory measures 15
1.4.1 Research question I: Measurement properties of memory self-reports across age 16
1.4.1.1 Measurement invariance in the confirmatory factor analytic approach 18
1.4.2 Research question II: Relation between memory self-reports and individual differences in memory performance measures – an emphasis on learning 21
1.4.2.1 Formalizing learning 23
1.4.2.2 Relating learning parameters to memory beliefs 25
1.4.3 Research question III: Alternative approaches to examine self-reports and memory performance – relating monitoring and learning 27
2 Empirical evaluation of three research questions 29
2.1 Measurement of cognitive failures 30
2.1.1 Introduction 30
2.1.2 Method 35
2.1.3 Results 38
2.1.4 Discussion 47
2.2 Individual differences in verbal learning in old age 54
2.2.1 An empirical analysis of verbal learning in old age 60
2.2.2 Emprirical findigs 62
2.2.3 Conclusions 71
2.3 Age differences in the underconfidence-with-practice effect 79
2.3.1 Introduction 79
2.3.2 Experiment 1 82
2.3.2.1 Method 83
2.3.2.2 Results 84

2.3.2.3 Discussion 90
2.3.3 Experiment 2 93
2.3.3.1 Method 93
2.3.3.2 Results 93
2.3.3.3 Discussion 101
2.3.4 Conclusion 103
3 General discussion 107
3.1 Summary and discussion of the results 108
3.1.1 Measurement invariance of the cognitive failures questionnaire across the adult lifespan 108
3.1.2 Verbal learning as an alternative memory measure 109
3.1.3 Monitoring and learning in young and old adults 112
3.2 Coda and outlook 114
References 118
Appendix 136

Figure and table legends

Table 2.1: Sample Characteristics 36
Table 2.2: Model Fit Indices for Single Group Models 40
Table 2.3: Factor loadings and explained variances of the CFA three-factor model for the whole sample 42
Table 2.4: Model fit Indices for Multiple-Groups Models of the three-factor model 44
Table 2.5: Descriptive statistics and sample correlations of cognitive variables and age……63
Table 2.6: Sequence of Estimated Models and Fit Statistics 65
Table 2.7: Means and standard deviations of JOLs and recalled words in Experiment 1 85
Table 2.8: Means and standard deviations of JOLs and recalled words across both experiments 94

Figure 1.1: Illustration of the hierarchical organization of meta-level and object-level 5
Figure 2.1: Factor means with associated 84 % confidence intervals (CI's) 47
Figure 2.2: Seven randomly selected model-based trajectories 67
Figure 2.3: Model SVLM2 71
Figure 2.4: Means of JOLs and recalled words 87
Figure 2.5: Difference between JOLs and recall performance 89
Figure 2.6: Means of JOLs and recalled words from Experiment 2 97
Figure 2.7: Difference between mean JOLs and mean recall performance. 100

1 Introduction

1.1 Aging and memory performance

Research on cognitive performance across the adult lifespan has highlighted a broad point of consensus, namely the decline in performance. Among different domains of cognition, memory has certainly been one of the best researched (e.g., Craik, 1977; Craik, Anderson, Kerr, & Li, 1995; Kausler, 1994; Verhaeghen, Marcoen, & Goossens, 1993). Experimental and psychometric findings indicate, on average, age-related decrements in the ability to learn and remember (for an overview see Horn & Hofer, 1992; Salthouse, 1991; Zacks, Hasher, & Li, 2000). The decline can be described by a curvilinear trajectory: After a rapid performance increase during adolescence and a plateau phase during early adulthood, decline sets in that becomes negatively accelerated after onset of very old age (Baltes & Lindenberger, 1997; Lindenberger & Baltes, 1995; Rönnlund, Nyberg, Bäckman, & Nilsson, 2005; Schaie, 2005). Hence, age-related decline in cognitive performance is smallest from age of 35-60 years (Martin & Zimprich, 2005), and increases after age 60 (Rönnlund et al., 2005), and is even more pronounced in 85+-year-olds (Baltes & Lindenberger, 1997).

Reasons for the onset of decline around the age of 60 years may be linked to decline in biological functions (Prull, Gabrieli, & Bunge, 1999), expressed, for example, by diminishing brain weight (e.g., Greenfield et al., 1967), reductions in hippocampal volume (Raz, Rodrigue, Head, Kennedy, & Acker, 2004), and decreasing integrity of the dopamine system (e.g., Antonini et al., 1993; Bäckman & Farde, 2004). Rönnlund and colleagues (2005), however, noted that reductions in biological functions may have an earlier onset already at the age of 20 or 30 years. A similarly early onset of decline has been observed in working memory and perceptual speed both in cross-sectional (Park et al., 2002) and in longitudinal samples (Schaie, 1994, 2005).

Not all types of memory seem to be affected by decline in the same manner or to the same extent, though (Zacks et al., 2000). A typical finding is that different domains of memory follow different patterns of decline across the adult lifespan (Horn & Hofer, 1992; Rönnlund et al., 2005; Salthouse, 1991; Schaie, 2005). That is, decrements are typically slight in implicit memory tasks (e.g., test performance in priming tasks) in which a stimulus that has been presented previously affects current behavior when presented again, often without the person realizing that the stimulus was encountered beforehand. In contrast, age-related losses

are substantial in explicit memory tasks (e.g., performance in free or cued recall in a memory test) as well as in working memory tasks (Grady & Craik, 2000). Especially newly learned material seems to be affected by decline resulting in poorer performance when measured in laboratory settings (Baddeley, 1990; Salthouse, 1991). Rönnlund and colleagues (2005), for example, have examined the trajectory of semantic and episodic memory functioning in ten age cohorts ranging from 35 to 80 years. They reported a differential aging pattern of episodic and semantic memory performance between ages of 60-80 years with substantially more decline for the episodic memory measure. The discrepancy between the findings of a late onset of decline in declarative memory and an early onset of decline for more basic functions may have several explanations.

The negative influence of reductions in basic functions may be counterbalanced by compensatory mechanisms both at the neural and at the psychological level. In fact, at the neural level McIntosh and colleagues (1999) found evidence in a fMRI study that older adults showed additional activations in the left prefrontal cortex during a perceptual memory task compared to younger adults. This can be interpreted as representing recruitment for compensatory purposes in the "aged brain" (e.g., Grady & Craik, 2000). At the psychological level a protective factor against decline in cognitive functions could be age-related increase in knowledge with advancing age. That is, knowledge about cognition and memory may be seen as part of crystallized intelligence which comprises accumulated knowledge and experiences an individual has made (cf. Cavanaugh & Blanchard-Fields, 2006; Horn & Cattell, 1966). Following Horn and Cattell's classification of intelligence, crystallized abilities are less prone to age-related decline compared to fluid abilities such as reasoning or processing speed.

An important aspect of the knowledge domain which might prove to be the basis for compensating mechanisms in old age is *metamemory*, that is, metamemory entails the knowledge people hold about their own memory. In this thesis I will focus on the relation between metamemory and memory performance across the adult lifespan because it probably represents the most important psychological mechanism to compensate age-related memory decrease.

1.2 Aging and metamemory

Interest in the individual knowledge about general memory functioning may be explained by the hope that memory performance can be enhanced by improving individuals'

fundamental knowledge about how the different memory domains work and how they are interconnected. To investigate this relation people are asked to provide self-reports about their own memory functioning. This approach to examining beliefs people hold about themselves is not new to scientific psychology (e.g., Hertzog & Dixon, 2005). In fact, it bears on the concept of introspection which, for some time, was considered the best way to examine psychological states and processes which defied external observation. But already early in time critique emerged on the validity of self-reports and it was pointed out that introspection would potentially be unreliable (Compte, quoted in James, 1890). As a consequence, introspection was marginalized in the later behavioristically based psychology and its methods were radically rejected (Dunlop, 1912; Watson, 1913). However, the concept was recast and refined and gained back some influence in psychology (Ericsson & Simon, 1980). In his monograph on consciousness and metacognition, Nelson (1996) argued that "introspective reports … can be related to other empirical observations and thereby can help investigators draw inferences about the participants' psychological processing" (p. 103). The quality and usefulness of measures based on introspection might be taken with a grain of salt, given that people may "be treated as *an imperfect measuring device of* [their] *own cognitions,* in which the individual's metacognitive monitoring is assumed to sometimes contain errors or distortions" (Nelson, 1996, p. 106).

The basic idea of introspectionism, that is, asking subjects about their own beliefs about a certain psychological process is still used in the investigation of metacognition, which can be defined as cognitions about cognition (Wellman, 1983). The focus is not on how the person carries out these processes, but rather on what the person knows and beliefs about these processes (Wellman, 1985). In other words, people think about thinking, and this self-awareness includes beliefs about how and how well their mental activities function (cf. Lovelace, 1990). To measure metamemory it is best framed as consisting of three elements: *Knowledge* about memory and memory functions, the *monitoring* of the current state of the memory system, and *beliefs about memory* (cf. Hertzog & Hultsch, 2000). The latter element, beliefs about memory, can be further subdivided into beliefs about one's own memory and beliefs about others' memory. In the present thesis I will focus mainly on monitoring and the beliefs about one's own memory functioning, the latter also having been termed self-referent memory beliefs and as such are purely subjective self-reports in terms of introspection

(Hertzog & Dixon, 2005). Note that in the remainder of this thesis I will subsume monitoring and self-referent memory beliefs under the heading of *memory self-reports*.

The main goal of this thesis is to examine the correspondence between memory self-reports and their behavioral counterpart, that is, memory performance, across the adult lifespan. The thesis is structured in three parts: In the first part I will present main findings concerning the relatedness between memory self-reports and objective memory performance. These findings are discussed in the light of a lifespan perspective and three open issues regarding measurement properties of metamemory, the relation between metamemory and memory performance, and monitoring processes in relation to learning are presented. In the second part, three studies are presented which aim at clarifying the aforementioned open issues. In the third part of the thesis, I will evaluate the three studies and give an outlook for future research in the domain of metamemory aging.

1.2.1 Memory self-reports

As defined in the previous section, memory self-reports comprise monitoring and self-referent memory beliefs. *Monitoring* is defined as a function which allows us to evaluate our cognitive processes. The assumption in models using a meta-level is one of a hierarchical system with at least two levels. In metamemory, the lowest or object-level is memory functioning in a given situation. As an example one might imagine that a person learns a word list to subsequently recall it. Through monitoring this person informs herself about how well this list is memorized and if she has to expend more time or cognitive resources to attain the desired level of mastery (Koriat, Sheffer, & Ma'ayan, 2002; Nelson & Dunlosky, 1991). The next level is the meta-level, which is both controlling and monitoring the object-level (cf. Metcalve & Shimamura, 1994), that is, monitoring informs the meta-level about the state the object-level is in and controlling processes inform the object-level what to do next. In short, the object- and the meta-level together with monitoring and controlling processes form a feedback loop which aims at reducing the discrepancy between a desired and the actual level of the object-level (e.g., memory performance). The desired level is conceptualized in a model within the meta-level. A schematic representation of the flow of information is given in Figure 1.1, which is a modified reproduction of Metcalfe and Shimamura's (1994) model of metacognition.

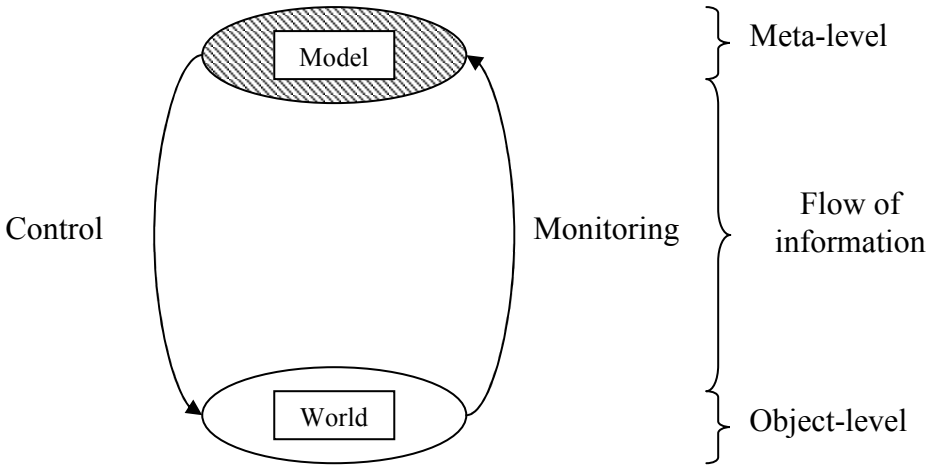

Figure 1.1: Illustration of the hierarchical organization of meta-level and object-level and the hypothesized flow of information (following the illustration by Metcalve & Shimamura, 1994).

Optimally, it seems highly desirable that monitoring does provide the exact or adequate information about the level of mastery. This would allow to achieve the desired level and to avoid unnecessary overlearning which costs time and cognitive resources. Hence, based on their metacognitive judgments individuals should be able to efficiently control and regulate their strategies for learning and retrieving information from memory (Schneider & Pressley, 1989).

The importance of a well-functioning memory monitoring may gain even more weight when memory performance is declining, as it is the case with older adults (cf. Kausler, 1994). The usefulness of monitoring is underlined by findings regarding the changes in memory and metacognitive monitoring across the lifespan which evidence that even though memory performance is impaired in the aging process, monitoring functioning appears to be spared from cognitive decline (Connor, Dunlosky, & Hertzog, 1997; Hertzog, Kidder, Powell-Moman, & Dunlosky, 2002; Shaw & Craik, 1989).

The second metacognitive concept discussed in this thesis is one of *self-referent memory beliefs*. Different from monitoring, which delivers information on-line about a current state, self-referent memory beliefs are available independently of a certain situation. A further important difference is that monitoring is best seen as a process which takes place for some time and which is specifically calibrated to a given task. Self-referent memory beliefs, in turn, represent beliefs about the state of one's own memory functioning in general and are

retrievable upon request. For example, if a person is asked about her general memory capacity or about the perceived change in memory performance in the last five years this person will access self-referent memory beliefs to answer the query. Other than monitoring, these beliefs do not appear to regulate memory functioning directly (but see McDonald-Miszczak, Hunter, & Hultsch, 1994) but may be seen as part of the person's self-concept (Silvia & Gendolla, 2001). Hence, memory beliefs are probably more prone to bias regarding self-consistency and more susceptible to incorporating implicit theories about general memory functioning than monitoring (Cavanaugh, Feldman, & Hertzog, 1998). Notwithstanding, these beliefs represent a source of information people rely on when estimating their memory functioning. Especially in the case of memory complaints, which become more pronounced as people age (Zimprich & Kliegel, in press), the "self-diagnosis" is usually based on self-referent memory beliefs. Again, a realistic appraisal of one's own memory functioning appears desirable because a diagnostic benefit could be drawn from these self-reports – given that they reflect the true state. That is, experienced changes would be related to true changes in memory performance. Furthermore, knowing its own strengths and weaknesses enhances the effective use of memory strategies which may be used to improve one's own memory functioning or to compensate memory decline due to deminishing cognitive resources in old age. One could argue that a realistic appraisal of one's own memory may support an optimal allocation of resources during a memory task and to the best possible strategy use, which, in turn, should lead to a maximization of memory performance. Consequently, the essence in both subjective measures discussed – monitoring and self-referent memory beliefs – is the degree of correspondence, that is, accuracy, between self-reports and objective memory performance.

1.2.1.1 Measurement of memory self-reports

In the measurement of monitoring three different meta-level statements can be obtained based on a distinction drawn by Nelson and Narens (1994): *Ease-of-learning* (EOL) judgments, which occur in advance of acquisition, are largely inferential and pertain to items that have not yet been learned. These judgments are predictions about what will be easy/difficult to learn, either in terms of which items will be easiest or in term of which strategies will make learning easiest. *Judgments-of-learning* (JOL) occur during or soon after acquisition and are predictions about future test performance on recently studied items. These recently studied items may be items for which there has not been a recall test or for which a

recall test occurred (irrespective of the correctness of the answer). *Feeling-of-knowing* (FOK) judgments occur during or after acquisition and are judgments about whether a given currently nonrecallable item is known and/or will be remembered on a subsequent retention test. When, for example JOLs are examined, respondents estimate the probability on a percent scale of recalling a learned item in a later memory test (e.g., Nelson, Dunlosky, Graf, & Narens, 1994; Perlmutter, 1978; Scheck & Nelson, 2005). To determine the degree of accuracy these predictions are then related to the actual recall performance in a subsequent memory test. Hereby, two conceptualizations of accuracy are typically calculated: *Calibration,* which refers to the correspondence between mean JOLs and mean actual performance, and *resolution*, which is commonly indexed by a within-participant Goodman–Kruskal gamma correlation (hereafter, gamma correlation) between JOLs and actual memory performance (Koriat, Ma'ayan, & Nussinson, 2006; Nelson, 1984; Nelson & Dunlosky, 1991).

In the case of self-referent memory beliefs verbal reports are usually measured by administering a memory questionnaire which inquires about different facets of memory comprising memory capacity, experienced memory change in a given period of time or experienced anxiety in a memory demanding situation and other domains. For example, an Item inquiring about experienced memory change might be formulated as follows “I’m less efficient at remembering things now than I used to be” (Item 14 from the Metamemory in Adulthood Quesionnaire (MIA); Dixon, Hultsch, & Hertzog, 1988). The verbal reports in form of sum scores from questionnaire data are then related to a test performance stemming from, for example, a free or associate recall memory test.

1.2.2 Age-related changes in memory self-reports

When inquiring about cognitive performance in general and about memory performance in particular, a predominant finding is that older adults report more negative beliefs and expectations compared to younger adults (Gilewski, Zelinski, & Schaie, 1990; McDonald-Miszczak, Hertzog, & Hultsch, 1995). This can either be due do experienced decline in memory performance or it can be due to a bias older adults’ hold about memory performance in old age – or both. Cavanaugh, Feldman, and Hertzog (1998), for example, argued that beliefs about memory may be seen as part of a common self-theory of aging: When asked about personal memory beliefs older adults are more likely to access memory-failure concepts and to make dispositional evaluations relative to young adults or to one’s

own past. That is, in old age, subjective judgments of one's own memory performance might be clouded by a general loss or decline expectancy. Similarly, McDonald-Miszczak, Hertzog, and Hultsch (1995) proposed a social-cognition framework, which posits that implicit knowledge about a general decline of cognitive functioning in old age might bias judgments of the elderly about their own cognitive functioning toward the general expectation of decline. Hence, memory beliefs change across the adult lifespan resulting in older people reporting more memory complaints compared to younger adults. However, how closely the reported decline truly reflects *individually* experienced decline in memory remains an open issue. McDonald-Miszczak and colleagues (1995) argued that the correspondence of memory beliefs and memory performance should increase into adulthood because older adults experience more memory failures, and hence, are more aware of the memory decline. It is important to note that age-related decline represented in memory self-reports is considerably smaller than change in memory performance. Given that self-referent memory beliefs show a less pronounced decline compared to actually measured memory performance suggests that the relation between the subjective and objective domain is not very strong. In fact, Devolder and Pressley (1991) compared perceptions about memory functioning in younger adults (mean age: 28 years) and older adults (mean age: 69 years). Whereas memory performance was better in the younger sample, perceptions about memory varied little as a function of age, and subjective memory was unrelated to objective memory performance in both age groups. Zelinski, Burnright, and Lane (2001) found that in a sample of 6,446 participants aged from 70 to 103 years the association between memory self-reports and memory was unrelated to age.

Other than self-referent memory beliefs memory monitoring may be less affected by implicit theories and expectations regarding its functioning because it is integrated in the feedback loop between the object- and the meta-level. Hence, one would argue that sources of bias become smaller as monitoring on a certain task advances leading to increasing accuracy (see also Chapter 2.3). Regarding its development the same question as for self-referent memory beliefs can be posed: Is the accuracy of memory monitoring subject to change across the adult lifespan? Results from two early studies suggest little age differences in the accuracy of monitoring as operationalized by JOLs. Lovelace and Marsh (1985) presented groups of young and old adults 60 high-frequency paired associates. Participants had to give JOLs of subsequent recall on a Likert-type scale. The authors concluded that there were no age

differences in memory monitoring accuracy during learning. Rabinowitz, Ackerman, Craik, and Hinchley (1982) studied monitoring accuracy by modifying the learning instruction. One half of young and old adults were given an intentional learning instruction whereas the other half received an imagery instruction. Further, JOLs were grouped into three levels (high, medium, and low confidence) and participants evaluated the probability of recall at each level. Older and younger adults given intentional learning instructions showed stronger relationships of JOLs to recall than did adults in the imagery instruction condition. More important, there were no age differences in these relationships which implied that monitoring was not related to chronological age. The methodological approach used in these early studies on monitoring accuracy has been criticized, though. The items presented in the experiments were not scaled in the same metric as the predictions elicited by the participants, and it is not clear what the conversion of ordinal Likert-type scale ratings to expected percentage of recall should be (see Hertzog & Hultsch, 2000). In a more recent study older participants reported different JOLs than young participants: Connor, Dunlosky, and Hertzog (1997) compared the metamemory accuracy of younger and older adults in a single experimental task. A mixed list of high- and low-association pairs was created to enable the evaluation of whether there were age differences in the sensitivity of predictions to the level of relatedness. Participants were able to differentiate high- and low-association pairs producing lower mean JOLs for the latter. At the same time older participants predicted, on average, higher levels of recall than younger adults did which led to a substantial overestimation of performance. In fact, the tendency for older adults, compared to younger, to overestimate their own performance proved to be a rather stable finding not only for monitoring but for metamemory judgments in general (Bruce, Coyne, & Botwinick, 1982; Mazzoni & Nelson, 1995; Schneider, Visé, Lockl, & Nelson, 2000).

1.3 Relation between memory self-reports and memory performance

Early research on the relation between subjectively assessed and objectively measured memory performance was based on the straightforward hypothesis which stated that there should be a strong association between memory self-reports and actual memory performance (Kail, 1990; Schneider, 1985). Thus, memory self-reports were expected to directly reflect actual memory performance – an assumption which resembles strongly the original idea of

introspectionism. In order to estimate the accuracy of memory self-reports, a simple approach is to compare the metamemory reports with the actual memory performance. In fact, this is a frequently used method to examine the relation between subjective and objective memory. More technically it has been formulated by Nelson (1996) who stated that:

> …the metacognitive approach is to formulate verbal reports as meta-level statements about what is occurring at the object-level, to operationalize what is occurring at the object-level through some kind of observable criterion response, and then to asses empirically the degree of relationship between the verbal report and the criterion response. (p. 106).

However, as has been demonstrated in a number of studies, the relation between self-referent memory beliefs and actual memory performance is moderate, at best (Arbuckle, Gold, & Andres, 1986; Barker, Carter, & Jones, 1994; Hänninen et al., 1994; Hertzog & Hultsch, 2000; Sunderland, Harris, & Baddeley, 1983). As pointed out in Chapter 1.2.2, on average, metamemory does not follow the same decline trajectory as memory performance implying that both measures are only weakly related. McDonald-Miszczak and colleagues (1995) tested this assumption in two samples comprising adults ranging in age between 22 and 86 years and concluded that the overall pattern resembled earlier findings where, despite significant changes in both memory performance and metamemory, the two sets of variables did not show a high degree of correspondence in change at either the mean or the individual level. In the majority of studies self-referent memory beliefs correlated between .2 and .3 with actual memory performance (Hertzog & Dixon, 2005; Niederehe & Yoder, 1989; Pearman & Storandt, 2004). The same holds for monitoring, where the correspondence between memory performance predictions and actually measured memory performance is rather low at first (Mazzoni & Nelson, 1995; Scheck, Meeter, & Nelson, 2004; Schneider et al., 2000). In a related vein of research, where self-referent memory beliefs are conceptualized as memory complaints, similar results have been reported: A consistent finding is that memory complaints are only weakly related to actual memory performance, indicating a lack of correspondence between memory complaints and actual memory performance (Derouesné, Lacomblez, Thibault, & LePoncin, 1999; Hertzog & Hultsch, 2000; Kliegel & Zimprich, 2005; Zimprich & Kliegel, in press). In fact, memory beliefs can be, from a correlational perspective, grouped with other self-related reports. For example, affective states and personality variables seem to play a major role in forming memory beliefs. A number of studies indicate that depressive and anxious symptomatology or other personality factors such

as neuroticism and conscientiousness may be much stronger related to self-referent memory beliefs than actual memory performance is (Hänninen et al., 1994; Niederehe & Yoder, 1989; Pearman & Storandt, 2004; Zimprich & Kliegel, in press). In a sample comprising 1,007 participants from the Interdisciplinary Study on Adult Development (ILSE; Martin, Grünendahl, & Martin, 2001), Zimprich and Kliegel reported a correlation between memory complaints and depressive affect of $r = .49$. One interpretation of this relation suggests that a person's affective state or personality colors the subjective evaluation of their cognitive performance, for example, with negative affect resulting in an amplification of subjective cognitive complaints (cf. Williams, Little, Scates, & Blockman, 1987). In turn, measures of memory performance fit in with other cognitive measures as, for example, processing speed, that is, those high in speed tend to recall more items in a memory test than those low in processing speed (Zimprich et al., in revision).

To summarize, the results from studies examining the relation between memory self-reports and objectively measured memory performance suggest two weakly related domains. In fact, the low correspondence between the subjective and the objective domain is not proprietary to the memory domain but is found in almost all sorts of self-reports, such as knowledge of attitudes (e.g., honesty in reports of self-relevant information), interoception of somatic states (e.g., perception of caffeine and alcohol symptoms), perception of self as a causal agent (e.g., assessing the own role in a given situation), and others (for more examples see Silvia & Gendolla, 2001). By and large, then one might agree with Wicklund and Eckert who stated that “people generally are not able to introspect accurately about their attitudes or other behavioral potentials” (1992, p.24).

1.3.1 Explanations for the low relation between memory self-reports and objective memory performance

In an attempt to explain the low correspondence between self-referent memory beliefs and actual memory performance and, at the same time, maintain the straightforward hypothesis of memory beliefs reflecting memory performance, Herrmann (1982) questioned the usefulness of questionnaires measuring subjective memory complaints. As later studies have shown, however, the measurement properties of questionnaires designated to measure self-referent memory beliefs are more than adequate, at least with respect to content validity, internal consistency, and factorial validity (Crook & Larrabee, 1990; Dixon et al., 1988). The

moderate correlation between memory beliefs and memory performance does not appear to be due to psychometric characteristics of questionnaires measuring self-referent memory beliefs. However, criterion validity which determines the degree of how well the questionnaire items capture the behavioral criterion, that is, memory performance, is rather low.

The psychometric properties of monitoring measures are less evident: A measure which best describes the degree of correspondence between JOL and recall performance is the within participant gamma-correlation (Nelson, 1984). Studies using this conceptualization of accuracy indicate that the relation between subjective and objective measure is moderate (Koriat, 1997; Koriat et al., 2002; Mazzoni & Nelson, 1995; Scheck et al., 2004; Schneider et al., 2000). However, if JOLs are repeatedly elicited in several study test trials the gamma-correlation increases steadily with every additional trial to high levels of correspondence. For example, Koriat (1997) administered in four study and recall trials 70 word pairs to a group of young university students. The gamma correlation after the first trial was .75 but steadily increased across the remaining three trials to .97. Another way to increase accuracy of JOLs is achieved if JOLs are given with a certain delay after the presentation of the stimulus. In that case the gamma-correlation increases markedly: Nelson and Dunlosky (1991) administered 60 test items to 33 university students. Half of the items received immediate JOLs, that is, right after the presentation of the item the subject was asked to rate the probability of recalling that specific item later and the other half received delayed JOLs. The gamma-correlation for immediately elicited JOLs was on average r = .36 but the gamma-correlation for delayed JOLs was on average r = .90. The authors termed this finding as the *delayed JOL effect.* Hence, the criterion validity of memory self-reports appears to be generally low at the first measurement occasion – but for monitoring, the measurement design can increase the correspondence between subjective and objective measures significantly.

Regarding validity, also the specificity of the self-referent memory beliefs questionnaires was investigated: Hertzog, Park, Morrell, and Martin (2000), for example, examined a cross-sectional sample of 121 adults, aged between 35-84 years, by administering questionnaires measuring depressive affect, memory complaints, and other variables, a set of cognitive tasks, and an interview about problems with remembering to take medications as prescribed. The highest correlations (ranging from r = .29 to r = .42) were found between subjectively reported failures in taking medications and the actually measured omissions or commissions concerning medication intake. The results of the study may be interpreted in

terms of behavioral specificity, implying that adult's self-reports of memory problems are valid when they focus directly on specific memory-related behaviors in everyday contexts. Thus, one way to increase the associations between self-referent memory beliefs and actual memory performance might be the examination of more specific domains of memory beliefs and to match these with corresponding memory tasks. Note, however, that the correlations increased only slightly and they are still smaller compared to the correlations between self-referent memory beliefs and other self reports as, for example, depressive affect.

Monitoring measures show per se a very high degree of behavioral specificity. Given that judgments are always elicited regarding a specific item, monitoring benefits already from higher correlations due to the specificity of the judgment. Hence, behavioral specificity is not an issue in monitoring measures.

In order to give an additional explanation for the lack of correspondence between memory beliefs and memory performance, researchers have begun to ask how subjects form self-referent memory beliefs. Commonly, memory beliefs questionnaires neither offer criteria for determining what constitutes memory problems or failures nor do they provide standards or anchors for scaling subjective memory problems (Hertzog et al., 2000). Respondents then might adopt very different criteria for what they consider weak or strong memory performance. Some persons might use a social comparison (cf. Festinger, 1954) by contrasting their performance with other people's performance, for example, people appearing brighter or less bright. Others might use an estimation of their own performance in earlier years, that is, a temporal comparison (cf. Albert, 1977). Still others might refer to how they performed in a particular, cognitively challenging situation (cf. Smith, Sala, Logie, & Maylor, 2000). Although preference of temporal comparisons appears to increase with age, all types of comparisons are existent in older adults (Brown & Middendorf, 1996; Robinson-Whelen & Kiecolt-Glaser, 1997; Suls & Mullen, 1983-1984).

For monitoring, the formation of JOLs has been discussed in three models, (1) in the cue-utilization framework, (2) in the anchoring-and-adjustment hypothesis, and (3) in the dual-factors hypothesis. Koriat (1997) proposed the *cue-utilization framework*, which distinguishes three types of cues for JOLs: Intrinsic, extrinsic and mnemonic. Intrinsic cues involve characteristics of the study items pertaining to its perceived difficulty (e.g., degree of associative relatedness between the members of a pair). Extrinsic cues relate to the conditions under which stimuli are learned and to the encoding operations applied by the learner (e.g.,

the number of times an item was studied and the amount of time the item was presented). The third type of cue comprises internal, mnemonic indicators that signal to the person the extent to which an item has been learned and will be recalled in the future (e.g., cue familiarity). According to the cue-utilization approach, in making JOLs participants do not monitor directly the strength of the memory trace of the item in question, but use a variety of cues that are generally predictive of subsequent memory performance. Because JOLs are based on inferences and heuristics, accuracy judgments depend on how the learner weights the importance of the cues for decision-making. A different perspective was adopted by Scheck and Nelson (2005). They hypothesized JOLs are resulting from an *anchoring-and-adjustment effect*. Briefly, anchoring may be described as a pervasive judgment bias in which decision makers are systematically influenced by arbitrary starting points (Chapman & Johnson, 1999). After having examined earlier studies where participants were required to judge recall probabilities for paired associates (Connor et al., 1997; Richards & Nelson, 2004), Scheck and Nelson (2005) concluded that a possible anchor is located around JOLs of 30% to 50%. Other than the cue-utilization approach, Scheck and Nelson assumed JOLs to be mostly unaffected by item difficulty, that is, no matter how easy or difficult items are, people tend to locate JOLs between 30% and 50%. In order to test their assumptions, the authors conducted an experiment with two learning and recall trials using easy and difficult items. The JOLs indeed seemed to be influenced by an anchor. Another, recently presented approach is the *dual-factors hypothesis*, which bears on both the anchoring and the cue-utilization approach (Scheck et al., 2004). The hypothesis states that the magnitude of JOLs derives both from an anchor point and from the on-line monitoring of items. The magnitude of JOLs is expected to change according to item difficulty, but not to the same extent as the corresponding recall level. The authors compared the adjustment process with the notion of the regression-toward-the-mean. In fact, JOLs given immediately after the presentation of paired associates resulted in large anchoring effects (Scheck et al., 2004). Conversely, delayed JOLs changed directly with item difficulty and were minimally affected by the anchor. Hence, for immediate JOLs the results were consistent with the anchoring hypothesis. In turn, delayed JOLs seemed to relate more on monitoring processes which means that participants rely more on cues pertaining to the ease with which the target can be retrieved than on an anchor.

In summarizing, the moderate correlations reported in most studies might underestimate the “true” relation between memory beliefs and actual memory performance

due to individually different comparison processes or criteria in forming memory beliefs. In line with this assumption, the longitudinal association between *changes* in memory beliefs and *changes* in memory performance is much stronger than the cross-sectional association, presumably because individual differences in criteria for forming memory beliefs are controlled for when people are examined longitudinally (Zimprich & Martin, 2002; Zimprich, Martin, & Kliegel, 2003). Hence, in order to enhance the relation between subjective and objective measures, one might formulate specific questionnaire items which closely correspond to the actual objective measure. For monitoring, one might delay JOLs or use repeated presentation of stimuli to enhance accuracy. Further, the use of longitudinal designs can help solve the problem of individually varying scaling criteria in memory questionnaires.

1.4 Open research questions on memory self-reports and memory measures

Based on the straightforward hypothesis that memory reports do reflect memory performance, a vast number of studies have been conducted to verify this relation (cf. Hertzog & Hultsch, 2000). The most replicated finding over the last years, however, was that metamemory and memory performance correspond only marginally (cf. Hertzog & Dixon, 2005; Nelson, 1996; Zimprich & Kliegel, in press). Up to this point, several factors have been identified which might be responsible for the low relation between self-reports and memory, such as low criterion validity and low specificity of memory questionnaires. Further, the social cognition framework and implicit theories about aging were developed to explain the lack of correspondence (Cavanaugh et al., 1998; McDonald-Miszczak et al., 1994). In the more specific domain of monitoring, several theories were developed to explain the formation of monitoring reports and to account for the lack of correspondence (Koriat, 1997; Scheck et al., 2004; Scheck & Nelson, 2005). These explanatory approaches, however, were never tested in a lifespan framework. Hence, little is known about age-related change in monitoring.

Most of the interest and effort to explain or overcome this lack of relation, was on the subjective measure, that is, memory questionnaires were criticized to be too general and more specific items about memory functioning were advocated (e.g., Hertzog et al., 2000). There was some success in reducing the low correlation between self-reports and objective memory performance, but the increase in correspondence hardly outweighs the loss in generalizability of the results. Hence, the advantage of behaviorally specific over general memory

questionnaires appears rather limited. In sum, one might agree with Wicklund and Eckert (1992) who stated that people are not able to report correctly about their attitudes or behavior.

Still, the straightforward hypothesis that memory reports do reflect memory performance is a valuable assumption which, in my opinion, has not been falsified entirely yet. In what follows, I will address three neglected topics in metamemory research which might help clarify the findings of the differential trajectories of memory-self reports and memory performance across the adult lifespan and I will offer a new approach by relating memory self-reports with memory performance.

1.4.1 Research question I: Measurement properties of memory self-reports across age

The first research question pertains to the measurement properties of self-reports. A common well documented finding is that memory self-reports reflect only moderately memory performance (Hertzog & Hultsch, 2000; Zimprich & Kliegel, in press). Age-related decline in self-reports appears to follow a different trajectory compared to age-related decline in memory performance (McDonald-Miszczak et al., 1995). Whereas there is consensus about the low relation between memory self-reports and memory, the investigation of age-related change in self-reports may be biased due to methodological artifacts. That is, self-referent memory beliefs are typically measured by questionnaires which were gauged on a certain age group or cohort. In order to examine age-related change in memory, a questionnaire might be administered to other age groups as well, for which the questionnaire has not been validated yet. Most of the times when different groups are examined researchers are interested in the comparison of a specific variable across these groups. Thus, an important precondition when comparing results from memory questionnaires across different groups, cross-sectionally or longitudinally, is the establishment of measurement invariance (MI) to demonstrate that a given test measures the same underlying factors across groups. The use of the expression "same factor" indicates that a factor has exactly the same conceptual interpretation across groups (cf. Lubke, Dolan, Kelderman, & Mellenbergh, 2003). For example, age differences in the structure of the MIA (co-)variances and MIA factor means can only be unambiguously examined if the measure of self-reports is unbiased with respect to age. Implicitly, most measurements of psychological constructs are conducted using tests assuming that the scores are an expression of an underlying latent trait or a common factor. When test scores are to be

compared across groups, it has to be ascertained that indicators (e.g., items of a memory questionnaire) of an underlying latent construct (e.g., memory capacity) mean the same thing to members of different groups. In many studies, however, it is implicitly assumed that the measures utilized to assess memory self-reports be invariant, an assumption that, if it goes untested, may lead to an over- or underestimation of age-related differences in memory-self reports. For example, in their research on adjectives selected to assess the Big Five personality factors (Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Culture) Zimprich, Allemand, and Huber (2007) identified an item which was not invariant across three age groups. More specifically, the adjectives "vulnerable – hardy" appeared to measure something different in the oldest group than in the two younger groups. The authors argued that a possible explanation might be that the bipolar adjectival marker "vulnerable - hardy" was interpreted from a more physically-oriented perspective by the old group, thus measuring rather subjective health, while in the young and middle-aged groups it was understood as it was intended, that is, as a description of emotional stability or morale. As a consequence, this specific item is not invariant with respect to the selection variable and a meaningful comparison across age groups can not be conducted. The items for which MI does not hold are said to be "biased" or to show "differential item functioning."

Hence, a main issue when comparing data across different age groups (or other selection variables as gender, cultural groups, racial groups, etc.) is one of accurate operationalisation. The operationalization calibrates manifest indicators to theoretical constructs, which are latent in the sense that they are not directly observed. Consequently, a measure is only interpretable and comparable across groups when it accurately and invariantly operationalizes the construct it purports to measure across the selection variable. The definition of MI states that, conditional on the factor scores, observed scores do not depend on group membership. This means that members of different groups who have the same score on the factor (e.g., the same level of ability) have on average the same observed scores. The definition of MI implies that groups may differ only with respect to the means and covariances of the factors that are measured by the observed scores (Lubke et al., 2003).

If the measures used so far in the investigation of memory self-reports prove to be measurement invariant with respect to age, the conclusions drawn to date are valid and the differential trajectory of self-reports and memory performance receives additional support. In

turn, if the measures are not invariant across age the conclusions drawn so far may be spurious and need to be revised.

In order address the issue of MI, I investigated the Cognitive Failures Questionnaire (CFQ; Broadbent, Cooper, FitzGerald, & Parkes, 1982) across six age groups comprising a total of 1,303 adults obtained from the Maastricht Aging study (MAAS; Jolles, Houx, van Boxtel, & Ponds, 1995). The objective was to establish strict measurement invariance which represents the strongest degree of MI across a large sample covering the adult lifespan in three factors, Forgetfulness, False Triggering, and Distractibility. The factor Forgetfulness comprises memory failures whereas False Triggering pertains to failures in cognitive or motor actions and Distractibility reflects mainly absentmindedness in social situations. The study is described and discussed in Chapter 2.1.

1.4.1.1 Measurement invariance in the confirmatory factor analytic approach

Given that MI is an issue of degree, which, borrowing from Meredith's (1993) terminology, ranges from configural invariance over weak and strong invariance to strict measurement invariance, it is important to distinguish between these different degrees. Some measures might be invariant at the configural level whereas others are strictly measurement invariant. Hence, when it comes to interpreting differences/similarities across the selection variable it is important to identify the degree of MI in order to draw adequate conclusions.

In this subsection, the different degrees of MI are discussed. Before I turn to the meaning of these degrees, the theoretical basis for MI is shortly revised. That is, MI is reviewed in the confirmatory factor analytic approach using continuous variables.

Mathematically it is possible to estimate the degree to which a measure is invariant regarding a certain selection variable. A formal definition of MI can be given as follows: Suppose a set of n measurements $\mathbf{Y}$, has been obtained on a random sample of subjects. Further suppose that these measurements are a statistical function of another set factor scores $\boldsymbol{\eta}$. Now consider a variable indicating group (here: age group) membership, denoted by $V=s$. The set of measurements $\mathbf{Y}$ is invariant with respect to s if

$$f(\mathbf{Y}|\boldsymbol{\eta}, s) = f(\mathbf{Y}|\boldsymbol{\eta}), \qquad (1.1)$$

where $\mathbf{Y}$ are observed scores and $\boldsymbol{\eta}$ are factor scores. Given a subjects factor scores $\boldsymbol{\eta}$, the subject's observed scores $\mathbf{Y}$ do not depend on group membership. This definition of MI has gained widespread consensus (see Meredith, 1993; Millsap & Everson, 1993). In practice, MI can be investigated by fitting multigroup confirmatory factor analysis (CFA) models to a given data set. To represent MI, certain model parameters are restricted to be equal across groups. Both, the restricted model and the less restricted model are fitted to the data. The models may be compared by means of a likelihood ratio tests which can provide evidence that MI is tenable across the selection variable (Lubke et al., 2003).

Assuming that one has applied the same items (or scales) measuring different memory constructs in different groups defined by a selection variable, for example, age, MI may be evaluated by examining invariance in factor loadings, latent intercepts, and residual variances by means of a confirmatory factor analysis of memory questionnaires across these groups. Examining different degrees of MI is, thus, accomplished by employing multiple group confirmatory factor models with increasingly severe across-group restrictions on parameters (Allemand, Zimprich, & Hertzog, 2007; Zimprich et al., 2007). Before turning to the different degrees of MI, the confirmatory factor analytic model is shortly revised.

The factor analytic model with non-zero means of manifest and latent variables is specified as follows in the multigroup factor model:

$$\mathbf{Y}_s = \boldsymbol{\tau}_s + \boldsymbol{\Lambda}_s^t \boldsymbol{\eta}_s + \boldsymbol{\delta}_s, \qquad (1.2)$$

where $\boldsymbol{\tau}$ represents a vector of intercepts, $\boldsymbol{\Lambda}$ is the factor loading matrix, $\boldsymbol{\eta}$ is the vector of scores on the latent variables, and $\boldsymbol{\delta}$ is the vector of residual terms. From this model the covariance structure can be derived:

$$\sum_s = \boldsymbol{\Lambda}_s \boldsymbol{\Psi}_s \boldsymbol{\Lambda}_s^t + \boldsymbol{\Theta}_s, \qquad (1.3)$$

which expresses the variance and covariance matrix in group *s* with $\boldsymbol{\Sigma}$ denoting the covariance matrix of the items, and $\boldsymbol{\Psi}$ representing the covariance matrix of the factor scores with the variances of the residuals expressed as $\boldsymbol{\Theta}$.

The mean structure is captured in the following equation:

$$\boldsymbol{\mu}_s = \boldsymbol{\tau}_s + \boldsymbol{\Lambda}_s \boldsymbol{\alpha}_s, \qquad (1.4)$$

with $\boldsymbol{\alpha}$ denoting the factor means in group s.

Meredith (1993) has elaborated the relation between MI and the multigroup factor model. More specifically, he has shown how the parameters of the multigroup model have to be restricted such that the model represents MI.

The most basic level of measurement invariance is *configural invariance* or pattern invariance, which requires that the same item must be an indicator of the same latent factor in each group (Horn & McArdle, 1992). This model implies that similar, but not identical, latent variables are present and, hence, the regression intercepts, $\boldsymbol{\tau}$ in Equation (1.4), have to be invariant in the *s* groups. Configural invariance suggests that the factors represent the same theoretical constructs, but these constructs can not necessarily be compared directly across groups because of possible inequalities of measurement (Bauer, 2005). If configural invariance is not supported, the evidence argues against even similar factor patterns across groups. In practice, the model of configural invariance serves as a baseline model against which more restrictive models are tested.

To achieve *weak invariance*, the factor loadings, $\boldsymbol{\Lambda}$ in Equations (1.3) and (1.4), are constrained to be equal across groups, but no other equality constraints are imposed. This model implies that the same latent variables are being measured across groups. If this level of invariance holds, an unambiguous comparison of the factor (co-)variance matrices is warranted. That is, differences in relationships among manifest variables are attributable to differences in relationships among latent variables. If weak invariance is not supported, one might consider identifying and deleting problematic indicators so that weak invariance is supported. Note, however, that altering the nature of a standardized scale might be also suscpetible to capitalization on chance characteristics of the observed samples (Byrne, Shavelson, & Muthén, 1989).

The next level is *strong invariance* which is specified by constraining factor loadings, $\boldsymbol{\Lambda}$ in Equations (1.3) and (1.4), and intercepts, $\boldsymbol{\tau}$ in Equation (1.4), to be equal across groups. Equal factor loadings imply that the measurement of the latent variables is the same in all groups. Further, the invariance in the intercepts allows evaluation of mean differences in

latent variables. Consequently, factor mean differences are scale invariant and interpretable, that is, all *s* groups have the same intervals and zero points (Lubke et al., 2003).

The model of *strict invariance* extends the previous model by invoking residual variances, Θ in Equation (1.3), to be equal in all *s* groups. The difference between strong and strict invariance concerns the formation of the variance structure in Σ. In the strong invariance case, group differences in the Σ-variance could be attributable to two sources, namely, to variances of the latent variables (i.e., group differences in the diagonals of Ψ) *or* to error variances (i.e., group differences in the diagonals of Θ). In the model of strict invariance, group differences in variances of manifest variables are *only* attributable to group differences in variances of the latent variables, since error variances are invariant across groups (Chen, Sousa, & West, 2005; Widaman & Reise, 1997).

Summarizing, MI is a matter of degree and, hence, the investigation of questionnaires inquiring about memory beliefs might not imply that factors are un- *or* ambiguously interpretable across a given selection variable. Instead, dependent on the degree of MI researchers are given a bandwidth of warranted interpretations of group differences.

1.4.2 Research question II: Relation between memory self-reports and individual differences in memory performance measures – an emphasis on learning

In the first section of this thesis, the methodological appropriateness of assessing memory beliefs by means of questionnaires was discussed. The majority of studies concerned with the problem of relatedness between self-reports and memory aimed at the subjective measure (Herrmann, 1982; Niederehe & Yoder, 1989; Zimprich & Kliegel, in press). For example, Hertzog and colleagues (2000) advocated a behavioral specificity approach, which, however, reduces generalizability with respect to the measured self-reports.

Another perspective is to ask if the "objective" memory-measure itself is an adequate behavioral counterpart of the typically used self-report questionnaires? Hence, I propose a different approach to explain the low association between memory beliefs (the emphasis here is mainly on self-referent memory beliefs) and memory performance: If the relation between beliefs about memory and their respective behavioral correlates is to be examined, it has to be ascertained that both measures are actually covering the same construct. As mentioned earlier, the most common way to assess the accuracy of memory self-reports is to compare

questionnaire data with the performance in a laboratory memory test which typically comprises one learning trial with subsequent free recall (e.g., Arbuckle et al., 1986) or cued recall (e.g., Hänninen et al., 1994). As pointed out in Chapter 1.3, the correspondence between both measures is only moderate even though there are some ways to increase the correlations between questionnaire data and memory performance. But, instead of formulating behaviorally specific items (Hertzog et al., 2000) to increase the correspondence between both measures, one might try to enhance the generalizability of the objective memory measure. Relating a single trial memory tests to questionnaires inquiring about general memory performance may be regarded as unbalanced: Only part of the acquisition process is covered by the objective measure, that is, only the recall performance after the first trial can be related to memory self-reports. But what happens "with the kind of memory that remains after a single study trial" (Nelson & Narens, 1994, p. 22)? If a single trial memory task is the behavioral correlate of the questionnaire items then, in turn, people ought to respond to the items, as if they were merely considering single trial memory experiences; this is, at least, arguable. In order to increase the overlap between subjective and objective measures it has to be clarified what lay adults mean by "memory" when they are answering memory belief questionnaires (Parr & Siegert, 1993). Considering naturalistic situations, a person's goal is to master a certain body of information, for example, a list of foreign language vocabulary or new text material (Nelson & Narens, 1994). Note that "master" does not necessarily imply that a person wants to reproduce all stimuli but maybe there are some specific items this person wants to recall whereas others are negligible. The desired level of mastery might be achieved after one presentation or, more probable, after a certain number of presentations and rehearsals (Nelson & Dunlosky, 1994). Hence, the whole acquisition process until the desired level of mastery is attained, may be apprehended by lay persons as constituting memory performance (Tulving, 2001). If people are inquired about their own memory performance, they might base their judgment on recall performance *and* the acquisition process. From this mastery-perspective, one crucial factor influencing self-referent memory beliefs is learning, apart from recall alone. Here, the amount of learning necessary is a crucial factor leading to the desired level of mastery. For example, imagine a participant in a longitudinal study lasting over ten years who always remembers the same amount of items in a word list: In the first measurement occasion this person needed three trials whereas ten years later five trials were required to remember the same amount of items. Hence, total memory performance remained

stable across the years, the effort to achieve the goal, however, increased over the same time course. Consequently, this person may report change in memory performance because more effort had to be expended to achieve the desired level of mastery, at the same time a recall test may indicate no change at all. Hence, relating a recall score obtained from a memory test to memory self-reports may miss an important part, namely the effort one had to invest in learning to achieve the recall level.

In summarizing, in everyday memory functioning, it seems unusual that people acquire information by single trials (Nelson & Dunlosky, 1994), therefore, the cognitive test may be adapted to that effect. The new approach here is to gain more information about the learning process without renouncing on the classical memory measure. To do so, an extension of the typical single-trial recall task to a learning task (e.g., a word list which is to be memorized during several trials) may represent a broader source of relations between the objective and the subjective measure compared to just one recall event. The main advantage is the possibility to model individual learning curves for each participant which can be related to self-referent memory beliefs. Hence, the correspondence between self-referent memory beliefs and memory performance is not just limited to one recall event, but now it is possible to further capture the learning rate and the performance after having administered a given number of trials.

1.4.2.1 Formalizing learning

In the following two subsections I will address the formalization of learning trajectories into learning parameters in order to relate these parameters to other variables of interest. The typical performance on a verbal learning task improves with repetition, but with every additional presentation the amount of performance improvement decreases. In consequence, the trajectory of recall performance follows a curvilinear form which constitutes the so-called learning curve (Ritter & Schooler, 2001). In order to relate learning to other variables of interest, for example, scores from memory belief questionnaires, it has to be made amenable to statistical analysis. In specifying the learning process by mathematical functions, a possibly large amount of variables and data points can be reduced to a few relevant parameters describing the learning process in a parsimonious way. These few parameters can easily be related to other variables of interest. Depending on the type of trajectory one wants to describe, different functions can be applied. That is, different

trajectories are best described by different functions. Apart from identifying the best fitting function, the approach of formalizing learning curves is the same as for any other curve. Hence, it is not limited to learning but can be applied to any repeatedly measured variable if certain preconditions are met.[1] Given that the process of learning meets all necessary conditions, it can be framed by mathematical functions describing nonlinear trajectories. Two basic types can be distinguished on the basis on how parameters enter the equation: Linear and nonlinear functions.

In polynomial functions all parameters enter the equation linearly and the curvature is achieved by adding or subtracting polynomials from each other or from constants. However, it has been recognized that low order polynomials do not always fit the data well. Higher order polynomials follow the data more closely but often fit badly at the extremes of the observed range of the axis of abscissae (Royston & Altman, 1994). A further disadvantage is that polynomials do not have asymptotes and can not fit when limiting behavior is expected (McCullagh & Nelder, 1989). For example, in the following function

$$f(t) = i + t - \gamma * t^2 \qquad (1.5)$$

where t denotes number of trials $t = 0,1, 2, \ldots t$ in a test, i is the intercept, and γ is a parameter which determines the shape of the curvature, all parameters enter the equation linearly. If this function is applied to a learning experiment, i typically represents the recall performance after the first trial. With every consecutive trial, the polynomials $t\text{-}\gamma*t^2$ adds (or subtracts) a portion from the intercept and the curve is shaped in dependence of γ. However, the second term, $\gamma*t^2$, exceeds t at some point with the consequence that the test performance is getting smaller again and eventually ends at $-\infty$. This follows from the global behavior of polynomials which are determined by their leading terms, here $-t^2$. Hence, with growing t, the function can be reduced to $-t^2$ which resembles an inverse U-form having both extremes at $-\infty$. Consequently, the usefulness of polynomials in modeling nonlinear trajectories is restricted to the number of observed trials the function is based on, and interpolation to a larger number of trials will *always* lead to wrong estimates.

[1] The trajectories must be transponable in functions of the Richards-family (Richards, 1959).

Functions, where parameters enter the equation nonlinearly do nut suffer from this behavior because boundaries can be set by upper or/and lower asymptotes. Consider, for example, the exponential growth curve

$$f(t) = \alpha + (\beta - \alpha)\exp(-\exp(\gamma)t), \qquad (1.6)$$

with the three parameters α, β, and γ. α denotes the upper asymptote, β, denotes the performance at the first trial and γ determines the growth rate of the trajectory. At the first trial, $t = 0$ and the function reduces to β, the initial performance. With growing t, the term $\exp(\exp(\gamma)t)$ approximates zero and at $\lim_{t\to\infty} f(t)$ the function reduces to α, the upper asymptote. Extrapolations to any given number of t are now possible because the function is limited by the upper asymptote and the unrealistic situation where values grow to infinity, as in polynomial functions, can be controlled and avoided.

If one considers nonlinear functions from a theoretical perspective, functions with asymptotic boundaries are compatible with, for example, verbal learning processes. That is, commonly verbal learning increases with every additional presentation of the stimuli but, as known for example from "testing the limits" studies (Baltes & Kliegl, 1992), after a certain number of trials the performance remains constant and increasing the number of trials will not improve performance. This behavior can not be modeled by polynomial functions because they grow to infinity.

1.4.2.2 Relating learning parameters to memory beliefs

Due to the aforementioned advantages, the focus in this thesis is on nonlinear functions. These can be recast in latent curve models which allow modelling both fixed effects and individual departures from these average effects (Blozis, 2004; Browne, 1993). Formally, an individual-specific approach requires expanding the fixed effects verbal learning curves as expressed in Equation (1.6) by random effects. If random effects are denoted by latin letters—as opposed to fixed effects, which are typically referred to by greek letters—the exponential learning curve of individual i becomes

$$f(t) = \{\alpha + a_i\} - (\{\alpha + a_i\} - \{\beta + b_i\})\exp(-\exp(\gamma + g_i)t), \quad (1.7)$$

which looks rather formidable. However, under standard assumptions about random effects (zero mean, normality), Equation (1.7) represents a model that belongs to a general class of estimable latent curve models (cf. Meredith & Tisak, 1990; Richards, 1959). Yet, to the best of my knowledge, growth curve-type models including random effects have not been employed in examining verbal learning curves. Part of the problem might be that learning curves usually are inherently nonlinear, which renders standard, that is, linear or quadratic, growth models inappropriate. Although nonlinear growth models have also been elaborated (see Cudeck & Harring, 2007; Davidian & Giltinan, 1995; Molenberghs & Verbeke, 2005, chap. 20), both the specification and estimation of such models become complex, especially so in large samples. As a viable alternative, Browne (1993; Browne & Du Toit, 1991) suggested to apply "structured latent curve models" for learning data, which impose specific nonlinear constraints on the pattern matrix of otherwise standard latent growth curve models. This more manageable and tractable approach was followed in the investigation described in Chapter 2.2 in order to model individual differences in verbal learning.

Note that focusing on individual learning curves instead of drawing mainly from grouped data is not an entirely new idea. Heathcote and colleagues (2000), for example, fitted nonlinear functions directly to individual data. Such an approach, however, holds the shortcoming that it requires a two-step procedure if inferences about population parameters shall be drawn: In a first step, individual learning curves are estimated, which, subsequently, in the second step, represent the raw data of further analyses, for example, in modelling the average learning curve. By contrast, the approach presented here allows for simultaneously estimating individual parameters *and* parameters that characterize the whole sample. This does not only have the advantage to result in correct standard errors of parameters estimates. Also, in estimating the parameters characterizing one specific individual, it allows for borrowing strength, that is, use information, from other individuals whose trajectory of verbal learning is similar (see Bryk & Raudenbush, 1992; Goldstein, 1995; Molenberghs & Verbeke, 2005). This is not to say that the description of average changes is invaluable. To the contrary, like it is common for mixed effects models, the present approach relies on the assumption that individual learning curves follow the same functional form as the average learning curve (cf.

Davidian & Giltinan, 1995). At the same time, modelling the average learning curve above all provides a means to an end to bring the individual back into learning curves.

In Chapter 2.2, this approach is demonstrated on the basis of a learning experiment with five trials. A number of models of verbal learning in old age which are based on different nonlinear functions are tested. The best fitting model is further extended by incorporating age and processing speed as explanatory variables.

1.4.3 Research question III: Alternative approaches to examine self-reports and memory performance – relating monitoring and learning

An alternative approach, which combines high item or stimulus specificity with verbal learning experiments is proposed in this section. Compared to the way self-referent memory beliefs are typically measured (i.e., by means of memory questionnaires), judgments of learning represent highly specific reports about the probability of recalling a given item in future. Every judgment is based on one specific item in a determinate situation, for example, the respondent is giving JOLs in a laboratory setting and is informed that recall will occur in ten minutes. Hence, studies examining monitoring are benefiting from the advantage in accuracy of behavioral specific settings over studies examining self-referent memory beliefs by means of questionnaire data (cf. Hertzog et al., 2000). In addition, the memory measure can be expanded from a one-trial recall task to a learning experiment. By supplementing the procedure with additional learning and recall trials a more naturalistic learning and monitoring situation is approached. Comparable to the learning process, monitoring is hardly limited to one single monitoring event. That is, monitoring is a process which is integrated in the feedback loop between object- and meta-level which, in naturalistic situations, is not limited to one single trial (Metcalve & Shimamura, 1994; Nelson, 1996).

In fact, the approach of administering more than one learning and recall trial is not new to the investigation of JOLs. Koriat (1997), for example, administered in four studies word pairs to a young group of university students across several study and recall cycles. In dependence of the metamemory accuracy conceptualization two rather different results emerged. While resolution or relative accuracy increased with every additional presentation of the stimuli, calibration or absolute accuracy did not increase steadily. That is, as memory performance increased with practice, absolute accuracy of JOLs decreased and became

underconfident after the second presentation. Koriat referred to this pattern as the underconfidence-with practice (UWP) effect. This effect has been replicated a number of times and it appeared to be robust against a number of experimental manipulations (Koriat et al., 2002).

There have been several studies examining the UWP effect and JOLs in young populations but, to the best of my knowledge, older adults have never been investigated with regard to the UWP effect. There was some effort to examine JOLs in older populations, though (e.g., see Connor et al., 1997). In these studies, older adults appeared to be overconfident regarding their recall performance The experimental settings for testing JOLs in older adults, however, were limited to one presentation and recall cycle only which does not allow drawing conclusions about the learning process end even less so about the UWP effect in older adults. Hence, little can be said about the developmental aspect of monitoring or more specifically about the UWP effect.

The investigation of monitoring in a learning setting in older adults is important for several reasons: Monitoring offers immediate information about the difficulty of a specific stimulus and it probably makes significant contributions to the success of learning, that is, monitoring informs the meta-level about the degree of mastery which, in turn, evokes more or less cognitive effort to attain the desired level. Given that monitoring is best seen as a process, it seems more fruitful to examine it in a multitrial context. This, however, has not happened with older adults yet, hence, almost no information is available about the evolvement and change in accuracy of monitoring in this group. As a further consequence nothing is known about the UWP effect in older adults. This effect might connote normal or successful memory functioning, that is, being underconfident keeps up the cognitive effort. With older adults being generally more prone to overconfidence in memory judgments, the UWP effect might be reduced which might also imply that older adults spend less effort on learning because they are too confident regarding their recall performance.

In order to investigate the trajectory of JOLs during a learning experiment and the UWP effect in older adults, two experiments were conducted with young and old adults. The experiments are reported in Chapter 2.3.

2 Empirical evaluation of three research questions

In this Chapter, three empirical approaches addressing the research questions are presented. First, a study on measurement invariance in memory self-reports utilizing the CFQ in a large sample is investigated. The aim is to establish strict measurement invariance in order to draw meaningful conclusions about age differences in the factor scores. Second, an approach to formalizing and relating learning to other variables of interest is presented. The main research question is addressed only indirectly by presenting the most appropriate technique to investigate data from learning experiments. Hence, the second study might be viewed as a demonstration of how learning and related variables of interest can be formalized using structured latent curve model. And third, JOLs and more specifically the UWP effect are examined and compared across young and old adults in order to gain more information about the development of monitoring into old age.

2.1 Measurement of cognitive failures[2]

2.1.1 Introduction

Slips and errors attract attention not only in everyday life - sometimes as comical mistakes such as putting flour in one's own coffee, or as more serious lapses such as turning the wrong way in a one-way street - but also in psychology (Broadbent et al., 1982; Reason, 1988; Wallace, 2004). A number of researchers have set out to examine the mechanisms underlying such everyday slips and failures, which are believed to originate from the cognitive organization of the individual. Norman (1981), for example, subdivided cognitively-based slips and failures into three categories of mistakes: errors in the formation of intentions, faulty activation of schemas, and false triggering of actions. By contrast, Reason (1988, 1990) attributed failures observed at the skill-based level of performance to two kinds of control deficits in automatic, noneffortful cognitive processing (cf. Hasher & Zacks, 1979): inattention and overattention. Inattention is considered to lead to capture slips, omissions following interruptions, reduced intentionality, perceptual confusions, and interference errors. In turn, overattention is regarded to result in omissions, repetitions, and reversals.

Another prominent account of everyday slips and errors is that of cognitive failures as proposed by Broadbent et al. (1982). A cognitive failure "… may involve perceptual failures, failures of memory, or physical actions which are misdirected. The common element is that there is a departure from the normal smooth flow of function, and events do not proceed in accordance with intention" (p.1). The assumption underlying cognitive failures is that various perceptual, action, and memory failures are influenced by a general and rather enduring factor, which might be described as a general proneness or liability to cognitive failures and which should be relatively independent of traditional personality and intelligence measures (cf. Klumb, 2001).

The Cognitive Failures Questionnaire (CFQ)

To assess the frequency of everyday cognitive failures, Broadbent et al. (1982) developed the Cognitive Failures Questionnaire (CFQ), which comprises 25 items derived from three areas of slips and errors: perception slips (e.g., fail to notice something relevant),

[2] I gratefully acknowledge the help of Daniel Zimprich, Martin Van Boxtel, and Jellemer Jolles in preparing the manuscript.

memory slips (e.g., absentmindedness), and slips in motor functioning (e.g., action slips). Respondents are offered examples such as "Do you fail to notice signposts on the road?", "Do you read something and find you haven't been thinking about it and must read it again?", "Do you bump into people?", and are asked to report the frequency of these incidents in the past six months on a five point Likert-type scale.

A number of studies have shown that cognitive failures, as measured by the CFQ, are related to, for example, absentmindedness (Reason & Lucas, 1984), slow performance on focused attention tasks (Meiran, Israeli, Levi, & Grafi, 1994), automobile accidents and work accidents (Larson & Merritt, 1991; Wallace & Vodanovich, 2003), dissociative experiences (Merckelbach, Muris, & Rassin, 1999), daytime sleepiness and boredom proneness (Wallace & Vodanovich, 2003), computing losses due to forgetting to save one's data in human-computer interaction (Jones & Martin, 2003), and reduced cognitive inhibition (Bloem & Schmuck, 1999). Regarding its psychometric properties, the CFQ has more than adequate test-retest reliability, with stability coefficients being around r_{tt} = .80 across six to 65 weeks, indicating a high degree of stability of individual differences (Broadbent et al., 1982; Merckelbach, Muris, Nijman, & de Jong, 1996; Vom Hofe, Mainemarre, & Vannier, 1998). The same authors provided coefficient alpha measures for the CFQ being around .90 although Merckelbach et al. (1996) reported somewhat lower alpha values in three samples, ranging from .75 to .81, implying more than adequate internal consistency. In an attempt to delineate a nomological network of cognitive failures, Wallace (2004) examined the associations between CFQ total scores and comparable constructs (e.g., neuroticism, absentmindedness, thought occurrence) as well as opposite constructs (e.g., conscientiousness, everyday attention, everyday memory, and action state orientation). In a sample of 386 undergraduate students he found that the frequency of self-reported cognitive failures correlated positively (*r*s = .50 to .53) with similar constructs, whereas the associations with opposite constructs were negative (*r*s = -.13 to -.41). The broad acceptance and usefulness of the CFQ are also reflected by the fact that the CFQ has been translated into several languages, for example, Dutch (Merckelbach et al., 1996), German (Klumb, 1995), Hebrew (Meiran et al., 1994) and Spanish (García Martínez & Sánchez-Cánovas, 1994). In summary, the CFQ is a commonly used questionnaire which has proved to be a useful instrument to identify individuals prone to cognitive failures.

Factor Structure of the CFQ

In most applied studies the sum score across all CFQ items is used as a measure of being prone to everyday slips and errors, based on the assumption that the CFQ captures a general liability of cognitive failures. In accordance with this assumption, Broadbent et al. (1982) conducted a number of factor analyses in different samples and concluded that a single, general factor of cognitive failures adequately captured the dimensional structure of the CFQ, because apart from the "obvious general factor" (p. 5), results were rather variable. Subsequently, however, several investigators re-examined the factor structure of the CFQ and their results seem to question the notion of only one single and general factor (Larson, Alderton, Neideffer, & Underhill, 1997; Pollina, Greene, Tunick, & Puckett, 1992; Wagle, Berrios, & Ho, 1999; Wallace, 2004; Wallace, Kass, & Stanny, 2002). Details regarding these models can be retrieved from Wallace (2004), where in the Appendix a tabular comparison of models is given. Thus, by contrast to the original surmise of Broadbent and colleagues, according to later findings the CFQ appears to be composed of more than one factor.

A number of researchers aimed at finding a more adequate dimensional representation of the CFQ by means of factor analysis. However, almost all researchers used principal components analysis (PCA), which represents a procedure to reduce data and may not be considered as the best approach to identify latent factors (Floyd & Widaman, 1995). Matthews, Coyle, and Kraig (1990), for example, administered the CFQ to a sample of 475 college students. They found two components, a general component and an additional component relating to memory for names, constituted only by two items, though. Larson et al. (1997) examined the structure of the CFQ in a sample of 2,379 American Navy recruits. By their own assertion, two components appeared to "incorporate a hodgepodge of different types of items" (p. 31) and, thus, were not meaningfully interpretable. In conclusion, the authors argued for a general component in terms of Broadbent et al. (1982) and a "memory for names"-component. In a recent study with 335 participants (223 undergraduate students and 112 US Navy personnel), Wallace and colleagues (2002) reported a solution that emerged from a PCA followed by varimax rotation, which yielded four components: Memory, Distractibility, Blunders, and Names. In a subsequent confirmatory factor analysis (CFA) in a sample of 709 university students, these findings were replicated (Wallace, 2004). Pollina et al. (1992) examined the structure of the CFQ in a sample of 387 college students. A PCA yielded five components: distractibility, misdirected actions, spatial/kinaesthetic memory,

interpersonal intelligence, and memory for names. Only three components, however, were considered reliable, of which distractibility alone accounted for 27% of the variance.

To summarize, with respect to the components underlying the CFQ, findings have been mixed: The structures of the presented solutions differed across authors both with respect to their content and complexity. Single-component to five-component solutions have been reported, but only few were replicable in independent samples. In fact, only the solution by Wallace et al. (2002) was retested and confirmed by means of CFA (Wallace, 2004). This heterogeneity in results may stem in part from the approach used to extract the alleged factors. By relying on PCA, the variance for discriminability of differences among possible factors is maximized, even more so, when varimax rotation is applied. Factors are forced to be independent which may not represent the factor structure best (Preacher & MacCallum, 2003). Hence, rather than factors representing dimensions of the CFQ, independent components were extracted which may have masked interfactor relationships and, as a consequence, may also have contributed to the diversity of solutions. Furthermore, investigations of the factor structure of the CFQ have been mainly based on young, adult populations. Consequently, it is unclear whether any of the previously presented solutions can be generalized---both in terms of the general structure and with respect to measurement properties--- to other populations. In retrospective, then, one might say that Broadbent et al.'s (1982) assertion that every sample yields a new factor structure seems to be the most stable finding over the years.

Cognitive Failures across the Lifespan

An underrepresented aspect in previous research on cognitive failures is whether the frequency of self-reported slips, errors, and lapses changes across the lifespan (but see Boomsma, 1998). There are, however, reasons to expect that the self-reported frequency of some cognitive failures increases into old age. Lay impressions hold that older adults are more forgetful, absentminded, and clumsy than younger adults (Heckhausen, Dixon, & Baltes, 1989) all of which are attributes that form part of cognitive failures. More generally, it was found that attributes carrying negative connotations, such as being indicative of memory failures or cognitive failures, are believed to be more pronounced in older persons, both by younger and older adults (Lineweaver & Hertzog, 1998). Consistent with these lay impressions, if older adults are asked to judge their own cognitive or memory functioning, usually a negative relation between age and self-reported cognitive or memory performance

emerges (Bolla, Lindgren, Bonaccorsy, & Bleecker, 1991; Derouesné et al., 1999; Hertzog, Hultsch, & Dixon, 1998). At the same time, individual differences in subjective assessments of one's own cognitive or memory functioning are only weakly related to individual differences in one's actual cognitive and memory performance as measured by psychological tests (Hertzog & Hultsch, 2000; Ponds, van Boxtel, & Jolles, 2000; Zimprich et al., 2003), implying that subjective judgments of cognitive functioning are, at best, partly based on objective performance. An explanatory account for these findings was offered by McDonald-Miszczak, Hertzog, and Hultsch (1995), who proposed a social-cognition framework which posits that implicit knowledge about a general decline of cognitive functioning in old age might bias judgments of older persons about their own cognitive functioning towards the general expectation of decline. That is, in old age, subjective judgments of one's own cognitive performance might be clouded by a general loss or decline expectancy. Similarly, Cavanaugh, Feldman, and Hertzog (1998) pointed out that memory failures may be seen as part of a common self-theory of aging: When asked about personal memory beliefs, older adults are more likely to access memory-failure concepts and to make dispositional evaluations relative to young adults or relative to one's own past.

Based on these arguments, one might hypothesize that the self-reported frequency of cognitive failures, especially those tapping failures associated with memory problems, is increasing into old age. Cognitive failures, although sometimes funny, carry a negative connotation and, as such, are believed to increase into old age, both by younger and older adults themselves. In addition, older adults are inclined to judge themselves based, in part, on a general cognitive loss or decline expectancy. Eventually, older persons appear to focus on failure episodes during the last six months instead of counterbalancing cognitive failures and success events. Together, these interrelated processes may lead to an increase in reported cognitive failures.

Aims of the Present Study

The aims of the present study were three-fold. First, we set out to find an adequate factorial representation of the CFQ in a large, representative sample covering the whole adult lifespan. To do so, we investigated previously reported factor solutions by means of confirmatory factor analysis. Additionally, we conducted an exploratory factor analysis followed by oblique rotation. Second, starting from a model based on an exploratory three-factor solution, we tested for different degrees of measurement invariance of the CFQ across

six age groups in order to examine whether the CFQ is unbiased with respect to age. The third aim was to, after having established strict measurement invariance of the CFQ across age groups, investigate age differences in factor covariances, variances, and means.

2.1.2 Method

Participants

The sample for this study comprised individuals from the Maastricht Aging Study (MAAS), a longitudinal study on the biological determinants and cognitive consequences of normal aging, stratified by age, sex and occupational achievement. In an early phase of MAAS, the sample was obtained through the registration network of family practices (RNH) supervised by the Department of General Practice of the University of Limburg. All participants were brain healthy; individuals with documented CNS pathology or MMSE scores below 24 were excluded (for a detailed description of inclusion criteria and sampling methodology refer to Jolles et al., 1995).[3] The main study of MAAS consisted of four cross-sectional panels, A1 to A4, sharing the same methodology with respect to sample frame, subject inclusion, stratification criteria, and basic measurement protocol. In the first wave of the MAAS study the CFQ was part of the assessment in three panels, A2 to A4, summing to 1,354 participants. In the present study, participants ranging in age from 24 to 83 years (M = 51.2, SD = 16.2) who had complete data records with respect to the CFQ were included. 51 participants (3.8% of the total sample) were excluded from further analyses as they did not provide complete data records concerning the CFQ, constituting a sample size of N = 1,303 participants, 49% of them female. Missingness of CFQ data was unrelated to age, gender, and educational level. The sample was split into six age groups, which, in the remainder of this study, will be referred to as Group 1 (Age: 24 – 33 years, M = 27.9, SD = 2.9) the reference group, Group 2 (Age: 34-43 years, M = 38.1, SD = 2.7), Group 3 (Age: 44-53 years, M = 47.7, SD = 2.6), Group 4 (Age: 54-63 years, M = 57.9, SD = 2.7), Group 5 (Age: 64-73 years, M = 67.8, SD = 2.7), and Group 6 (Age: 74-83 years, M = 76.3, SD = 2.1) (for descriptive statistics see Table 2.1).

[3] A detailed description regarding the rationale, design and methods of the MAAS can be retrieved at the project homepage:
http://www-np.unimaas.nl/maas/Moreinfo/MAAS_PB_intro.pdf

Table 2.1: *Sample Characteristics*

		Age Groups						
		1	2	3	4	5	6	Total
	N	227	232	237	228	229	150	1303
Age	*Mean*	27.98	38.05	47.74	57.95	67.82	76.31	51.21
	SD	2.86	2.66	2.55	2.74	2.68	2.09	16.22
Gender	% female	48.0	51.3	49.4	47.4	48.9	48.0	48.9
Educational Level[a]	*Mean*	4.66	4.26	3.81	3.08	2.92	3.16	3.68
	SD	1.67	1.65	1.86	1.68	1.77	1.98	1.87

[a]Measured on a scale ranging from 1 = primary education to 8 = university education, based on the Dutch educational system.

Across the six age groups, there were no differences in the proportion of female participants ($\chi 2 = 0.88$, $df = 5$, $p > .97$). Age groups, however, differed significantly in level of formal education ($F = 35.83$, $df = 5, 1297$, $p < .01$), indicating that, on average, younger age groups were better educated. According to Cohen's standards (Cohen, 1988), this effect was of medium size and explained 12% of total variance in education.

Measures

At first measurement occasion in 1994/1995, part of the data collection protocol of MAAS was the Dutch version of the Cognitive Failures Questionnaire (CFQ; Broadbent et al., 1982), a 25 item self-report inventory tapping different aspects of cognitive failures.[4] For each item, participants were asked to assess the frequency of a specific cognitive failure event they had experienced over the last six months using a five-point Likert-type scale. The scale-points of the 25 items are anchored by the descriptors *very often* (assigned the value 4) through *never* (assigned the value 0). The internal consistency (Cronbach's alpha) in the present sample was $\alpha = .89$, which is comparable to earlier studies. Note, however, that the internal consistency measure implies a unidimensional structure, which might represent an inappropriate assumption for the CFQ. This cautionary note is substantiated by the moderate

[4] The English version of CFQ can be found at http://www.atkinson.yorku.ca/~psyctest/cogfail.pdf.

mean polychoric interitem correlation of $r = .35$. Due to the Likert-type scale response format, the observed variables were treated as ordered-categorical in all subsequent analyses. Because of a very low answer prevalence in the fifth answer category ('*very often*') and its complete absence in some of the 25 items in some age groups, this category was collapsed with the fourth category in order to make it amenable to the analysis of measurement invariance using Mplus (Muthén & Muthén, 2004), resulting in possible total scores ranging between 0 and 75.[5] A detailed description of the confirmatory factor analysis model of ordered-categorical variables is given in the Appendix.

Statistical Analyses

In all subsequent analyses we treated the items of the CFQ as ordered categorical and used WLSM to estimate confirmatory as well as exploratory factor solutions. Statistical modelling proceeded considering a sequence of nested confirmatory factor models based on previous findings. First, the general factor model by Broadbent et al. (1982), the two-factor models by Larson et al. (1997) and Mathews et al. (1990), the four-factor model by Wallace (2004), and the five-factor model by Pollina et al. (1992) were estimated in the MAAS sample. In addition to the confirmatory factor analyses (CFA), an exploratory factor analysis (EFA) followed by oblique rotation was conducted in order to find an adequate and interpretable dimensional representation of the CFQ in the MAAS sample. A three-factor exploratory solution was, subsequently, re-estimated as a confirmatory model after having fixed non-significant factor loadings smaller than .15 to zero. Note that, contrary to earlier approaches, we used EFA and not PCA to find dimensions underlying the CFQ. In fact, EFA is considered widely as the appropriate approach for identifying dimensional structures underlying psychological constructs whereas PCA may be seen as a data reduction procedure (see Floyd & Widaman, 1995).

The model of three correlated factors was subsequently tested for increasing levels of measurement invariance across six age groups (Meredith, 1993; Millsap & Yun-Tein, 2004; Widaman & Reise, 1997). Measurement invariance was investigated as a series of nested models of first order factor solutions (Martin & Zimprich, 2005; Meredith, 1993; Vandenberg, 2002; Zimprich, Allemand, & Hornung, 2006)..

[5] Analyses including all 5 categories led to essentially the same results with respect to the analyses based on the total sample.

All analyses were conducted using Mplus, version 3.0 (Muthén & Muthén, 2004), applying the WLSM estimator.[6] As criteria for *absolute* model fit, the Root Mean Square Error of Approximation (RMSEA) and the incremental Comparative Fit Index (CFI) are reported. Values of the CFI above .90 are considered to be adequate and values above .95 indicate close model fit, whereas for the RMSEA values less than .08 indicate adequate model fit (Browne & Cudeck, 1993). Moreover, goodness of fit was evaluated using a rescaled χ^2-test, namely, the T_2^*-statistic proposed by Yuan and Bentler (2000), because data did depart from the multivariate normal distribution. In comparing the *relative* fit of nested models, ΔT_2^*-differences were tested for statistical significance utilizing the procedure described by Satorra and Bentler (2001). Note that, due to its dependency on sample size, the ΔT_2^*-difference test provides rather high power for large sample sizes. We therefore complemented it by calculating the CFI difference. As Cheung and Rensvold (Cheung & Rensvold) have demonstrated, if ΔCFI between two nested confirmatory factor models is smaller or equal to .01, the null hypothesis of equal fit of the two models should not be rejected. One has to keep in mind, however, that the critical values recommended by Cheung and Rensvold are based on a simulation study using maximum likelihood estimation in two groups, whereas we used the WLSM estimator in an ordered categorical sample with six groups, hence, this criterion may not perfectly fit to our situation. Still, although not explicitly suited for confirmatory factor models of ordered-categorical variables, we chose the ΔCFI as the main criterion due to its independence of sample size (cf. Marsh, Balla, & Hau, 1996).

2.1.3 Results

Dimensionality of the CFQ

Confirmatory factor analyses of the ordered categorical CFQ items started with the one-factor model arrived at by Broadbent et al. (1982). As indexed by the CFI and the RMSEA this general factor model evinced an acceptable absolute fit (see Table 2.2). On

[6] Apart from the WLSM estimator, which represents a mean adjusted Weighted Least Square estimator, Mplus also offers a mean- and variance- adjusted estimator (WLSMV). The WLSMV estimator does not allow for difference testing, however, because the degrees of freedom may vary within a given model specification. Furthermore, Asparouhov (2005) has demonstrated that WLSM and WLSMV performed equally well in medium to moderately large samples and both clearly outperformed WLS.

average, the single factor explained 30% of variance, ranging from 15% (Item 4) to 43% (Item 17).

Subsequently, the two-factor model reported by Larson et al. (1997; cf. Matthews et al., 1990) was tested. In terms of absolute model fit, the two factor solution yielded virtually identical results as the one-factor model (see Table 2.2), with the RMSEA and the CFI indicating acceptable fit. As the two-factor solution was nested in the one-factor solution, model fit comparison based on the CFI was warranted. Compared to the one-factor model, the two-factor model did not represent a critical improvement in relative model fit due to the ΔCFI value which was not exceeded (see Table 2.2). The two factors were strongly correlated (r = .77) and, on average, the model explained 31% of item variance, ranging from 15% (Item 4) to 49% (Item 7). Due to the result from the relative fit index, this model was not judged to fit data substantially better than Broadbent's (1982) solution. Given that both models were statistically not distinguishable, the more parsimonious unidimensional model was maintained. Note that, compared to the analyses reported by authors using PCA, here, factors were allowed to correlate.

Afterwards, the four-factor model suggested by Wallace (2004) was examined. Similar to the preceding models, absolute model fit was adequate with respect to the RMSEA and the CFI (see Table 2.2). The relative fit index of the Wallace model, which was nested in the previous models, did not indicate neither a critical difference to the unidimensional model nor to the two-factor model because the ΔCFI did not exceed .01. The four factors were strongly correlated (mean interfactor correlation r = .84), indicating almost a collapsing of factors, that is, factors that, in the present sample, were not separable. On average, 33% of the variance was explained, ranging from 16% (Item 4) to 48% (Item 7). Given that the criterion for an increase in model fit was not met, this model was not judged to fit the data better than the unidimensional model.

Next, the five-factor model of the CFQ arrived at by Pollina et al. (1992), which has a noncongeneric structure because Item 14 and Item 18 each load on two factors, was tested. Noncongenerity is given when at least one item is allowed to load on more than one factor, contrary to a congeneric structure where each item is associated with one factor only. Note that noncongenerity does not necessarily affect nestedness of models, in fact, this five-factor solution was still nested in the previous models. Absolute model fit indicated that the hypothesized model captured the observed data adequately (see Table 2.2).

Table 2.2: *Model Fit Indices for Single Group Models*

Model	No. of Factors	T_2^*	df	CFI	RMSEA	ΔT_2^*	Δdf	ΔCFI
Broadbent et al.	1	2066.05*	275	.950	.071	-	-	-
Larson/ Matthews	2	2003.42*	274	.952	.070	46.14*	1	.002
Wallace	4	1816.64*	269	.957	.066	156.12*	5	.005
						203.97*[a]	6	.007[a]
Pollina et al.	5	1747.54*	263	.959	.066	60.60*	6	.002
						257.95*[a]	12	.009[a]
EFA Three-Factor Model	3	1168.21*	228	.974	.056	528.75*	35	.015
						796.32*[a]	47	.024[a]
CFA Three-Factor Model	3	1293.35*	257	.971	.056	127.04*	29	.003
						549.31*[a]	18	.021[a]

Note. [a]Compared to Broadbent et al.; T_2^* = rescaled Chi-Square statistic; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; ΔT_2^* = difference between two rescaled T_2^* -statistics, calculated according to Satorra and Bentler (2001); Δ*df* = difference in degrees of freedom; ΔCFI = difference in CFI. *p< .01.

Compared to Broadbent et al.'s (1982) unidimensional model, the relative model fit ΔCFI did not denote a better fit for the five-factor solution. Furthermore, it did not outperform Wallace's (2004) four-factor model either. Factor inter-correlations were, in general, rather strong (mean interfactor correlation was r = .81). The five factor model explained 34% of the variance, ranging from 18% (Item 4) to 48% (Item 7). Again, the criterion for an

improvement of model fit was not met, therefore the solution of Pollina and colleagues was not considered to fit the data better than Broadbent et al.'s or any of the other models.

To summarize, the fit of models reported previously in the literature met the recommended cut-off criteria for the CFI and the RMSEA for adequate model fit. At the same time, each model in the sequence improved slightly, but significantly, in absolute model fit indices. At first glance, this might imply that a multidimensional solution seems to more adequately capture the structure of the CFQ. However, in terms of both absolute fit indices, CFI and RMSEA, and the amount of explained variance, differences in fit between models were, at best, marginal. Furthermore, taking into account the ΔCFI-criterion, the difference across the four tested models never exceeded the cut-off value of .01, indicating no critical differences in model fit. Eventually, factor intercorrelations in multiple factor models were, partly, inflated, which made it almost impossible to separate them as different dimensions of the CFQ. Taking these findings into consideration, none of the conceptually rather different solutions clearly outperformed the original model by Broadbent et al. (1982) which was still the most parsimonious solution among the tested models.

In order to arrive at a more consistent dimensional representation of the CFQ in the present sample, we conducted an EFA followed by an oblique promax rotation, with the number of factors ranging from one to five. With respect to both absolute and relative fit and in terms of interpretability of the factors, a model of three intercorrelated factors represented the data best (see Table 2.2, EFA three-factor model). The exploratory solution was re-estimated using confirmatory factor analysis (CFA) to obtain a more parsimonious solution. For the confirmatory analyses, only significant factor loadings ($p < .05$) were maintained in the model. Hence all factor loadings yielded by the exploratory analysis smaller than 0.15 in absolute value were set to zero, represented by empty cells in Table 2.3. The confirmatory three-factor model evinced a good fit, as indexed by the CFI and the RMSEA (see Table 2.2, CFA three-factor model). Note that this noncongeneric three-factor solution was nested in the solutions reviewed earlier. Compared to the one-factor model proposed by Broadbent et al. (1982), the confirmatory three-factor solution led to a substantively meaningful increment in relative model fit because the ΔCFI (.021) exceeded the critical value of .01. Factor 1, which was defined by high loadings of Items 1, 2, 5, 7, 17, 20, 22, and 23, may be interpreted as signifying "Forgetfulness," that is, a tendency to let go from one's mind something known or planned, for example, names, intentions, appointments, and words.

Table 2.3: *Factor loadings and explained variances of the CFA three-factor model for the whole sample*

CFQ Item	Forgetfulness	Distractibility	False Triggering	R^2
1	0.422		0.359	0.33
2	0.612	-0.506	0.892	0.48
5	–0.298	0.383	0.677	0.41
6	0.234		0.576	0.36
7	0.360	0.393		0.33
13	0.265		0.537	0.35
16	0.276	0.483		0.34
17	0.518		0.502	0.46
20	1.112		–0.268	0.49
21	0.230	0.586		0.37
22	1.031			0.52
23	0.389		0.544	0.41
8		0.575		0.25
9		0.624		0.28
10		0.566		0.24
11		0.679		0.32
14		0.722		0.34
15		0.483	0.299	0.27
18		0.328	0.598	0.44
19		0.798		0.39
25		0.635		0.29
24			0.656	0.30
12			0.881	0.44
3			0.829	0.41
4			0.519	0.21
Factor Correlations				
Forgetfulness	1.00			
Distractibility	0.74	1.00		
False Triggering	0.62	0.77	1.00	

Note. Only significant factor loadings ($p < .05$) are reported. Factor loadings smaller than 0.15 were set to zero.

Factor 2, which incorporated Items 8, 9, 10, 11, 14, 19, 21, and 25, reflected "Distractibility," mainly in social situations or interactions with other people, such as being absentminded or easily disturbed in one's focused attention. Factor 3, which comprised high loadings on Items 2, 3, 5, 6, 12, 18, 23 and 24, mirrored "False Triggering," that is, interrupted processing of sequences of cognitive and motor actions. In sum, the three factors explained 36% of the total variance in the sample. Factor 1 correlated with Factor 2 (r = .74) and Factor 3 (r = .62). The correlation between Factor 2 and Factor 3 was slightly higher (r = .77). Due to its improved fit, we decided to examine different degrees of measurement invariance for the three-factor model.

Measurement Invariance Across Age

The baseline model, configural invariance, requires that the same item must be an indicator of the same latent factor in each group hereby factor loadings can differ across groups. This model yielded an acceptable absolute fit (see Table 2.4), implying that configural invariance of the CFQ holds across the six age groups. Next, factor loadings were constrained to be equal across groups to test for weak invariance. According to the absolute fit indices, the model represented the data adequately, with the CFI remaining stable whilst the RMSEA improved to some degree. Relative model fit did not show a practically important difference to the preceding model because the ΔCFI did not exceed .01. In sum, one might conclude that weak invariance of the CFQ holds across the six age groups. In the following model, thresholds of the 25 items were constrained to be equal across groups to obtain strong invariance. As indexed by the CFI and the RMSEA, the fit of the strong invariance model was adequately capturing the data. The relative fit index, ΔCFI, did not indicate a change of substantive interest in fit compared to the weak invariance model. On balance, fit indices suggested that strong invariance of the CFQ holds across the six age groups. Next, strict measurement invariance was obtained by constraining residual variances to be equal across all age groups. Again the absolute model fit indicated adequate fit with a stable CFI and slight improvement in the RMSEA. The relative fit index did not denote a practical difference to the preceding model.

In summary, we concluded that there were no important differences in the relevant parameters between the six age groups across the configural throughout the strictly invariant model (see Table 2.4).

Table 2.4: *Model fit Indices for Multiple-Groups Models of the three-factor model*

Model	T_2^*	df	CFI	RMSEA	ΔT_2^*	Δdf	Δ CFI
Configural Invariance	3332.73*	1587	.957	.071	-	-	-
Weak Invariance	3530.03*	1742	.957	.068	249.62*	155	.000
Strong Invariance	3723.48*	1962	.957	.064	284.06*	220	.000
Strict Invariance	3817.92*	2087	.958	.062	182.51*	125	.001
Strict MI, φ_{21}, φ_{31}, φ_{32}	4043.39*	2102	.952	.065	41.51*[a]	15	.006[a]
Strict MI, φ_{11}, φ_{22}, φ_{33}	3996.13*	2102	.954	.064	49.94*[a]	15	.004[a]
Strict MI, $\boldsymbol{\mu}$	4252.85*	2102	.947	.069	102.34*[a]	15	.011[a]
Strict MI, $\boldsymbol{\Phi}$	4119.75*	2117	.951	.066	81.81*[a]	30	.007[a]
Strict MI, $\boldsymbol{\Phi}$, $\boldsymbol{\mu}$	4470.83*	2132	.943	.071	157.77*[a]	45	.015[a]

Note. [a]compared to the Strict Invariance Model; T_2^* = rescaled Chi-Square statistic; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; ΔT_2^* = difference between two rescaled T_2^*-statistics, calculated according to Satorra and Bentler (2001); Δ*df* = difference in degrees of freedom; ΔCFI = difference in CFI. *p< .01

Considering the general guidelines by Cheung and Rensvold (2002) and the small fluctuation in the RMSEA, strict measurement invariance for the first order factors for the CFQ can thus be assumed to hold, implying that a comparison of factor (co)-variances and factor means across the six age groups is unbiased.

Age Differences in Cognitive Failures

First, age differences in factor covariances were compared across groups. To do so, the covariances between Forgetfulness, Distractibility, and False Triggering were constrained to be equal across the six age groups. Doing so did not lead to a substantively important decrement in absolute or relative model fit (see Table 2.4; Model Strict MI, φ_{21}, φ_{31}, φ_{32}).

Hence, one might conclude that the associations between the three factors Forgetfulness, Distractibility, and False Triggering are of the same magnitude in all six age groups.

Subsequently, to further investigate age invariance in measurement of cognitive failures, variances were held constant in each factor. Analyses again started from the strictly measurement invariant model. The absolute and relative fit indices did not yield a substantially worse model fit compared to the strict measurement invariant model (see Table 4; Model Strict MI, φ_{11}, φ_{22}, φ_{33}). Consequently, Forgetfulness, Distractibility, and False Triggering variances were interpreted as being stable across the present sample.

The next step was to constrain factor means to be equal across all age groups. We started again from the strict measurement invariant model: In this case, however, model fit indices deteriorated as a result to the constraints imposed (see Table 2.4; Model Strict MI, $\boldsymbol{\mu}$). Notably, the CFI value dropped below .95, which, at the same time, led to a substantial increment in the ΔCFI. In fact, the critical value of .01 was exceeded, indicating that this model fitted the data worse compared to the strict invariant model. As a result, the means in Forgetfulness, Distractibility, and False Triggering can not be regarded as being equal across the six age groups.

Next, the models with equal covariances and equal variances were combined to a single model which evinced an excellent fit and, compared to the strict measurement invariant model, did not lead to a substantial decrement in model fit (see Table 2.4; Model Strict MI, $\boldsymbol{\Phi}$). According to the ΔCFI criterion, this model was not distinguishable from the strict invariance model indicating that variances and covariances remained equal across all age groups. Note that equal variances and equal covariances necessarily imply equal correlations between factors across age groups. These correlations were r = .74 (Forgetfulness and Distractibility), r = .62 (Forgetfulness and False Triggering), and r = .77 (Distractibility and False Triggering).

Finally, all covariances, variances and means were constrained to be equal. As a result, model fit indices deteriorated substantially suggesting that factor means need to be freely estimated in order to avoid misfit (see Table 2.4; Model Strict MI, $\boldsymbol{\Phi}$, $\boldsymbol{\mu}$).

It thus seemed appropriate to further investigate factor means in the six age groups. In order to do so, 84% confidence intervals (CI's) were calculated, based on the model of strict measurement invariance and equal covariances and variances across age groups. Non-overlapping 84% CI's indicate that factor means are significantly different at the $p < .05$ level.

In turn, if the 84% CI in one age group overlaps with the 84% CI of another group, factor means are not significantly different at the $p < .05$ level (cf. Goldstein & Healy, 1995; Tryon, 2001). Given that equal factor variances can be assumed to hold across the age groups, differences in factor means can be readily interpreted as effect sizes in term of Cohen's (1988) standard: 0.2 stands for a small, 0.5 stands for a medium, and 0.8 for a large effect between the factor means of a given age group and the reference group, that is, Group 1. For ease of interpretation, the factor means in the reference group were set to zero. Figure 2.1, which is split in three panels, shows factor means and 84% CIs of the three factors Forgetfulness, Distractibility, and False Triggering across age. For example, the factor mean of Forgetfulness in Group 3 is 0.394, with its 84% CI ranging from 0.256 to 0.533, whereas the 84% CI of Group 1 ranges from -0.136 to 0.136. Because the two CI's do not overlap, Group 3 differs significantly from Group 1 in Forgetfulness, implying that participants in Group 3 rate themselves, on average, as more forgetful than participants in the youngest, the reference group.

In terms of statistical significance, the general picture that emerged with respect to means in the cognitive failures domain was: (I) Forgetfulness followed a roughly linearly increasing trajectory, implying that older persons rated themselves, on average, as more forgetful than younger adults. Group 1, the reference group, differed significantly from all other groups. In addition, Group 2 (mean: 0.275) showed a significantly lower mean than Groups 4 through 6 (means: 0.598, 0.629, 0.840). Eventually, the Forgetfulness mean in Group 3 (mean: 0.394) was smaller than in Group 6. Accordingly, effect sizes ranged from small (Group 1 versus Group 2: $d = 0.275$) to large (Group 1 versus Group 6: $d = 0.840$). (II) Distractibility means tended to remain stable in the first four age groups, followed by a decrease in the two oldest groups. Groups 1, 2 (mean: 0.086), and 3 (mean: 0.100) differed significantly from those of Groups 5 (mean: -0.294) and 6 (-0.310). Effect sizes were small (Group 1 vs. Group 5: $d = 0.294$) to medium (Group 3 vs. Group 6: $d = 0.410$). (III) False Triggering did not show a pronounced age trend and all factor mean differences were statistically non-significant. Effect sizes were all in the small range, with the difference between Group 1 and Group 3 ($d = 0.190$) marking the largest effect.

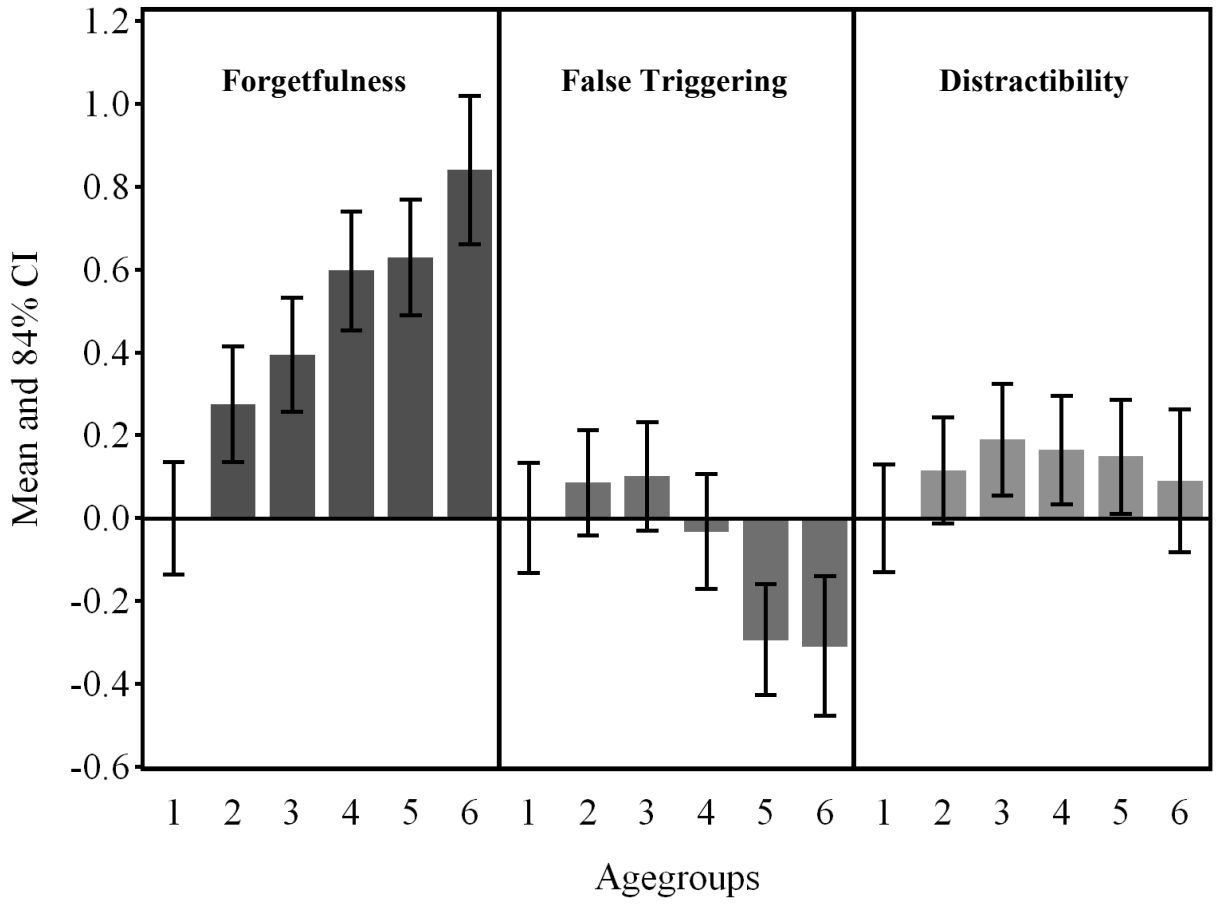

Figure 2.1: Factor means with associated 84 % confidence intervals (CI's) based on the strict invariant three-factor model with equal (co-)variances. Group 1 is the reference group with the factor mean 0. CI's, within a panel, not overlapping with the CI of the reference group indicate statistically significant differences at the 5% level. The left panel represents Forgetfulness, the middle panel represents Distractibility, and the right panel represents False Triggering. Due to equal variances, the factor means can be read directly as effect sizes, following Cohen's (1988) standard: 0.2 stands for a small, 0.5 stands for a medium, and 0.8 for a large effect between the factor means of a given age group and the reference group, i.e., Group 1.

2.1.4 Discussion

The present study pursued three aims. The first aim was to find an adequate dimensional representation of the CFQ in the sample of the Maastricht Aging Study (MAAS). The second aim was to test this solution for different degrees of measurement invariance to eventually, as a third aim, compare age-effects in the factors underlying the questionnaire.

A special feature of this study was the treatment of the CFQ items as ordered-categorical, which has not been taken into consideration in previous examinations of the CFQ. It has been shown that when analyzing Likert-scaled data as if they were continuous two

types of errors might occur: (1) Categorization errors from cutting continuous data into ordered categories and (2) transformation errors resulting from categories of unequal widths (O'Brien, 1985). Both types of errors attenuate the estimated relation between the latent construct(s) and observed items (DiStefano, 2002), even more so in the multiple-groups case (Lubke & Muthén, 2004). Considering the skewed distribution of the answer patterns in the items, which lead to the collapsing of the answer category "very often" with "often", and the multiple-groups analyses used in this study, treating ordered categorical variables as if they were continuous would have considerably biased the standard errors of parameter estimates in the factor analyses and distorted χ^2 -tests of measurement invariance (Finney & DiStefano, 2006). By treating Likert-scale data as ordered-categorical, an important source for parameter estimate bias can be minimized.

To investigate the dimensional representation of the CFQ in the MAAS sample, previously presented models were analyzed by means of confirmatory factor analysis (Broadbent et al., 1982; Larson et al., 1997; Matthews et al., 1990; Pollina et al., 1992; Wallace, 2004). For all tested solutions the fit of the models indicated a good representation of the data. However, it was found that previously reported multifactor models were afflicted by very strong factor correlations, which made it almost impossible to separate different dimensions of the CFQ. Furthermore, from a conceptual perspective one might consider previously reported multifactor solutions as unbalanced, because the number of indicators per factor/component is highly variable. This disproportion becomes obvious when considering, for example, the Larson et al. (1997) and Matthews et al. (1990) solutions, which consist of two components, one comprising 23 items and the other two items. In balance, none of the solutions previously reported in the literature managed to clearly outperform the other models with respect to model fit and distinctness.

The CFA solution we presented was derived from an EFA followed by oblique factor rotation. Note that re-estimating a factor structure by means of CFA from an EFA solution in the same sample may appear unwarranted, because, strictly speaking, doing so would usually require two samples, one for the calibration of parameters and the other for the cross-validation. However, as Floyd and Widaman (1995) pointed out, confirmation of an EFA solution will most likely fail if sample size is small and/or if the EFA solution fails to account for most of the systematic variance in the data. Both conditions were absent in our analyses. What's even more, in our opinion the fact that the same model showed strict measurement

invariance across six samples of different age (see below) may serve as providing sufficient support against spurious factor structures (see Reise, Widaman, & Pugh, 1993). The three resulting factors were interpreted as representing Forgetfulness, Distractibility, and False Triggering. In view of the fact that each of these three factors have emerged in previous studies (Meiran et al., 1994; Pollina et al., 1992; Wallace et al., 2002), they might be tentatively considered as being inherent to the CFQ. The oblique factor rotation and the noncongeneric structure of the three-factor solution lead to an attenuation of interfactor correlations up to a point where the three dimensions of the CFQ were distinguishable from one another. Furthermore, the noncongeneric structure of the model lead to a balanced solution regarding the number of items per factor which, at the same time, appears to imply that not all individual items of the CFQ are factor-pure in the sense that they measure one underlying latent variable only. That is, some items tap more than one factor, which one would expect, when considering the intertwinement of the three cognitive failure domains. To illustrate, see Item 2 of the CFQ: "Do you find you forget why you went from one part of the house to the other?" A respondent might agree to the item because he simply forgot his task or because he was distracted with something else and consequently could not remember why he went to the other part of the house or, a stimulus may have triggered another intention and the respondent subsequently ended up in the cellar instead of the washing room. Accordingly the Item can be associated to different domains of cognitive failures. On average, 36% of the total variance in the 25 items was explained. Although this might not seem too impressive, one has to take into consideration that factor analysis was conducted on the item level, where unsystematic influences tend to be more pronounced than in sum scores, where they tend to cancel out. Moreover, compared to previous analyses of the CFQ, our accepted model explained a relatively strong proportion of variance in the individual items. Still, however, this does not rule out the possibility that some systematic influences remained unaccounted for, for example, method effects like item wordings (Zimprich, Perren, & Hornung, 2005). Although Pollina et al. (1992) presented a noncongeneric solution as well, and Wallace (2004) allowed factors to be obliquely rotated, the combination of both, as presented here, has not been examined earlier.

The second aim of the study was to examine the measurement properties of the three-factor model of the CFQ. Specifically, we aimed at ensuring that the CFQ behaves equivalently across different age groups, that is, is free from age-related measurement bias.

For this purpose, measurement invariance (MI) across groups was tested in a sequence of four different hierarchical levels (cf. Meredith, 1993), which ultimately yielded strict MI to hold for the three-factor solution across the six age groups. Conceptually, establishing MI indicates that the meaning of Forgetfulness, Distractibility, and False Triggering is similarly comprehended by subjects throughout the six age groups. Taking into account the severity of restrictions that are consecutively imposed on the model, and the fact that an interpretable three factor solution was obtained, the finding of strict MI with respect to the CFQ across six age groups appears remarkable. Moreover, the size of the MAAS-sample implies considerably large statistical power (cf. MacCallum, Browne, & Sugawara, 1996). Furthermore, the good representation of the data justified the implementation of a noncongeneric model for the sake of a stable and well fitting and measurement invariant solution. This point of view is supported by Meredith and Horn (2001) who argued that, with regard to measurement properties, measurement invariance ought to be taking precedence over meta-theory (such as that of congeneric simple structure).

Instead of applying confirmatory factor analysis (CFA) of ordered categorical variables, as a viable alternative we might have used item response theory (IRT) models. As outlined by Reise, Widaman, and Pugh (1993), in principle, it is possible to examine different degrees of measurement invariance across groups by specifying according IRT models (see also Meade & Lautenschlager, 2004). However, from a practical side of view, there are some drawbacks in utilizing IRT models: First, setting up a model for a set of data using IRT seems much more advanced and less user-friendly---and, hence, more error-prone--- than CFA models. Second, the χ^2 measure of model fit results as a function of the difference between observed and expected response proportions, whereas the associated degrees of freedom are a function of the number of different response patterns minus the number of parameters estimated. As a consequence, model fits are not directly comparable between CFA and IRT models. Third, in IRT modeling, the only standard measure of fit is a likelihood ratio χ^2-variate, which is highly dependent on the sample size, whereas goodness of fit indices known from the SEM tradition have not been widely developed yet. Considering that goodness of fit indices represent the main basis to accept or reject a model, their absence in the IRT approach limited its applicability for our study.

The third aim of the study was to investigate age-related differences in covariances, variances, and means of the three factors. As strict measurement invariance across age held,

group differences in the three factors were meaningfully and unambiguously interpretable as reflecting only quantitative shifts in invariant measures. First, the covariance patterns of the three cognitive failure factors were compared across the age groups: Constraining covariances to be equal did not lead to a substantial decrement in model fit. This implies that there was no indication of any practically important age difference in the associations among the three factors. The association strength between Forgetfulness, Distractibility and False Triggering may tentatively be seen as remaining stable across the lifespan. Next, constraining factor variances to be equal across age groups did not lead to a relevant change in model fit. This finding implies that the amount of interindividual variability in the three factors was constant across the six age groups. Note that, the equality of factor or "true" variances and strict MI, that is, equality of "error" variances, in addition implies equal reliabilities of the manifest indicators across the six age groups (cf. Bollen, 1989). Due to the cross-sectional nature of the data analyzed in the present study, however, strong conclusions about perfect cognitive failure variance stability across the lifespan are to be drawn with caution. One ramification of age-invariant factor covariances and age-invariant factor variances is, however, that *correlations* among the three CFQ factors were also equal across the six age groups. This is a comparatively strong finding which implies that the structure of the three factors is scale invariant, that is, insensitive to change in scaling of the CFQ factors (Cudeck, 1989; Swaminathan & Algina, 1978). Eventually, factor means were constrained to be equal across the six age groups, which lead to a relevant decrement in model fit. The most apparent age-effect was observed for the Forgetfulness Factor, where a roughly linear trajectory of means indicated increasing self-reported Forgetfulness for older participants. Notably, the increase of Forgetfulness-means between the youngest and the oldest groups was large in terms of effect size. This finding is consistent with results from studies examining metamemory across the adulthood, where the relation between self-reported memory performance and age is negative (Bolla et al., 1991; Derouesné et al., 1999; Hertzog et al., 1998). Also, this result provides support for the assumption of implicit theories about aging and cognitive decline (McDonald-Miszczak et al., 1995) and the self-theory of aging (Cavanaugh et al., 1998), which predict an increase in reported memory complaints for older persons. The means of the Distractibility factor followed a different pattern: the mean response remained relatively stable across the first four age groups, comprising adults from 24 to 61 years. The two oldest groups, however, reported significantly less Distractibility than the younger groups. An explanation for the

sudden decrease might be that Distractibility is interacting with environmental factors, that is, factors not originating within a person as age-related, but as social or age-graded changes. Considering the sudden drop in the Distractibility mean, beginning in the early sixties, might suggest a linkage to a normative event, such as retirement from the job. Items loading on Distractibility like, "Do you leave important letters unanswered for days?" or "Do you find you forget appointments" might be answered by a retired person with "Rarely", simply because she has more time "to do things that work had precluded" (Nuttman-Schwartz, 2004, p. 235) compared to a person highly involved in work life or, she might answer the questions referring to their duties at work, which, after retirement, are not pertinent anymore. Some people might feel less distracted after retirement, because daily demands decrease in their number and hence the plentitude of tasks to be accomplished during the day diminish after retirement (Gall, Evans, & Howard, 1997; Quick & Moen, 1998). The third factor, False Triggering, did not show any significant mean-level changes across the six age groups. In addition, effect sizes were all marginal to small, which suggests that False Triggering taps a domain of cognitive failures that remains relatively stable across the lifespan. This finding is surprising because False Triggering may be seen as resulting from loss of activation in attentional resources (cf. Norman & Shallice, 1986). Lower levels of attentional resources for older persons have been documented in different research fields, for example, visual attention (Bedard et al., 2006), and dual task performance (Riby, Perfect, & Stollery, 2004). Norman (1981), however, remarked that subjects identify their cognitive failures only when they recognize a mismatch between their intentions and actions. Therefore, respondents may not regard their triggering errors as failures. Alternatively, the absence of an age effect in those items measuring False Triggering might also be due to the fact that they describe cognitive failures for which an increase across age is not expected by lay persons. Hence, even with implicit theories about aging being present in older persons, it might be that for some, possibly less frequent or less salient cognitive failures, age stereotypes are less clear-cut. Altogether, these findings highlight the diversity of cognitive failures, and importantly, they identify differential developmental trajectories of these three domains across the lifespan.

In conclusion, this study provides further evidence that the CFQ assesses multiple dimensions of cognitive failures, which proved to be strictly invariant over age groups comprising the adult lifespan. At the same time, strict MI with respect to age allows for extrapolations to other selection variables, because it almost certainly implies weak

measurement invariance for all selection variables correlated to age, for example, health status (Lubke et al., 2003). Whereas the three-factor model remains to be replicated across different samples, more work is needed to validate the three factors by relating them to similar constructs, for example, absentmindedness (Reason & Lucas, 1984), self-referent memory beliefs and memory complaints (Hertzog & Hultsch, 2000). If the factor solution presented in this paper proves to be stable, a potentially fruitful direction for future research is the investigation of age-related change in the three factors, as suggested by the Forgetfulness and Distractibility means. Note that an approach with one general factor only, as suggested by Broadbent et al. (1982), would have disguised age differences in cognitive failures because the underlying dimensions proved to be changing in opposite directions (Forgetfulness & Distractibility). As the three factors show, self-perception of cognitive failures is not a unitary system, but a composition of different dimensions changing in different rates and following different patterns of change over the life course.

2.2 Individual differences in verbal learning in old age[7]

Not too long ago, the psychological study of how people learn and remember verbal material has kept a number of eminent thinkers, scholars, researchers, and students occupied. Beginning in the 1950s, a variety of rather formalized statistical models of verbal learning emerged that were doing relatively well in capturing empirical data, mostly average learning curves (e.g., Bush & Mosteller, 1955; Estes, 1950). Despite their success, the work of these scientists by and large now stands there, unread, gathering dust on the shelves of many university libraries. The main reason for this might be that most of these models were formulated in a Stimulus-Response framework, which went out of style at the end of the 1960s and was replaced by the concepts and vocabulary of information processing. With this replacement, also the term"learning" lost much of the popularity it had in conjunction with verbal memory phenomena, while, at the same time, "memory" became the more often used notion (Nelson & Narens, 1994). However, because verbal learning necessarily entails acquisition, storage, and retrieval, verbal learning and memory research are, in fact, so closely connected that any distinction between these two parts of a bipartite field might be considered arbitrary. As Tulving and Madigan (1970) pointed out already at the beginning of the cognitive era of psychology, verbal learning and memory research might be described as two intertwined subcultures that share a common goal, but talk different languages and use different methods. Similarly, Craik (1977, p. 385) has argued that research into memory mainly utilizes once-presented material and one single recall trial, whereas examining verbal learning usually requires multiple presentation of material and several or multitrial recall cycles. With a grain of salt, then, one might assert that verbal learning captures the "dynamic" aspects of memory, that is, systematic changes in verbal memory performance due to repeated practice.

Taking up such a working distinction between verbal learning and memory, one has to diagnose that the bulk of research on verbal memory phenomena today is conducted using single recall trials, that is, it represents memory research. This holds also and is especially true for the investigation and comparison of memory performance in different age groups. A glimpse into the references section of Kausler's (1994) benchmark monograph on learning

[7] A similar version of this chapter was submitted for publication elsewhere (see Zimprich, Rast, & Martin, in press)

and memory in older adults shows that the majority of research on age differences in verbal learning, that is, multitrial free recall, was performed before 1980. And although we do not present any exact numbers here, we suspect that this situation has not changed very much since the publication of Kausler's book. This is to say that the interest in examining verbal learning in younger and older adults - as opposed to investigating their verbal memory -, appears to have become minimal, apart from, for example, issues in diagnosing dementia (Schoenberg et al., 2006). Such an unbalanced situation might not be without reasons (see above), but appears unwarranted in light of the fact that many naturalistic learning situations do not only comprise one study cycle, but rather involve several trials until a desired level of mastery is reached (cf. Nelson & Narens, 1994). This is even more true with respect to older adults, where the importance of learning for maintaining cognitive performance has been frequently stressed as providing an enormous preventive potential (e.g., Hultsch, Hertzog, Dixon, & Small, 1998; Martin & Zimprich, 2005; Stern, 2002; Willis & Schaie, 2005) and is becoming more and more of an issue in cognitive aging research (e.g., Willis et al., 2006).

Facing this state of affairs, we felt it timely to revive the interest in verbal learning in old age. Such an effort should, of course, not be interpreted as discrediting memory research in the elderly, but rather to complement and enrich it by taking a closer look on individual differences in learning in old age. However, our approach to verbal learning in old age differs from previous ones that have demonstrated that older adults show decrements in verbal learning (cf. Kausler, 1994). Instead of focusing on group-based data and, thus, the average learning curve, our goal was to reinstate the individual into the learning curve. More specifically, we aimed at modelling individual differences in learning in old age. Such an individual-centered perspective represents a fundamental shift from and extension of "traditional" verbal learning research, because it is less focused on the question of why older people learn at all, but rather asks why different older people learn differentially, that is, why older people differ in their amount and rate of verbal learning. Note that a similar shift from group-based to indiviudal-specific approaches has taken place in developmental aging research, where, during the last ten years, a number of studies began to emerge that provided new insights into cognitive aging by taking into account individual differences in change (e.g., Hofer & Sliwinski, 2006; Martin & Zimprich, 2005; Zimprich, 2002; Zimprich & Martin, 2002). After having thus clarified the setting and aims of our research, to begin with we introduce some formal models of the learning curve.

Representing learning in old age by means of nonlinear functions

Typically, performance on a task improves with repetition. However, with every repetition the amount of performance improvement decreases. In consequence of these two constituents, performance improvement or learning of a task may be described as a process that benefits from investing in further practice, but with diminishing returns. If performance is diagrammed as a function of the number of practice repetitions, the so-called learning curve emerges that follows a gradually increasing, albeit negatively accelerated trajectory (cf. Ritter & Schooler, 2001). The relation between performance improvement and repeated practice as described in the learning curve is so ubiquitous that it applies to a broad variety of performance increments in human behavior, for example, acquisition of new skills (e.g., Ackerman, 1988), gaining knowledge of statistics (e.g., Smith, 1998), and, of course, verbal learning (Tulving, 1964).

As noted above, learning curves describe the change in performance over trials t (t = 1,…, n). More formally, if learning is monotonically increasing, learning curves can be described using the following equation:

$$f(t) = \alpha - (\alpha - \beta) \cdot g(t) \qquad (2.1)$$

where α is the upper asymptote of the curve and β is the initial value of performance.[8] These two parameters act as boundaries, because the lower performance limit is given by β and the upper performance limit is given by α. The function *g(t)* describes the type of curvature present in the learning curve across the n trials. As such, *g(t)* might be called the core of the learning curve and is usually a function of a third parameter, a learning rate parameter γ (cf. Paul, 1994). A psychological interpretation of the three parameters is straightforward. The parameter β represents performance after the first trial, that is, after the first learning cycle is finished. Thus, it may be interpreted as initial performance level that closely resembles the performance that is measured using typical, one-trial memory tasks (cf. Hultsch et al., 1998). The presence of an upper asymptote (α) in Equation (2.1) suggests that learning tasks have

[8] As an aside, we note that counterexamples to a smooth, monotonic, concave-upward function predicted by this general model are abundant. If learning, for example, occurs in bursts of insight, there may be "jumps" in the according learning curve, a phenomenon typically occurring in problem solving (e.g., Jones, 2003). Also, for example, learning curves for copying Morse code often contain plateaus, where little progress is made, only to be followed by new increases in learning rate with further practice (e.g., Wisher, Sabol, & Kern, 1995).

natural ceilings or limits on performance. These asymptotes may be determined by the experimenter's choice of task material, for example, list length in free recall. Note that the asymptote is not necessarily reached within a given range of trials, but rather, as a limiting value, should be interpreted as potential maximum performance, that is, a prediction of a subject's performance after an infinite amount of training (cf. Browne, 1993; Browne & Du Toit, 1991; Mazur & Hastie, 1978; cf. Meredith & Tisak, 1990; Richards, 1959). Eventually, the learning rate γ denotes the rate of approach from initial level to potential maximum performance. Larger values of γ correspond to faster rates of learning, that is, higher quantums of improvement in performance.

As candidates for *g(t)*, different authors have suggested different core functions. For example, Heathcote, Brown, and Mewhort (2000) have advocated the exponential curve, the core function of which is given as $g_{ex}(t) = \exp(-(t-1)\gamma_{ex})$, which, after substituting into Equation (2.1) leads to

$$f_{ex}(t) = \alpha_{ex} - (\alpha_{ex} - \beta_{ex})\exp(-(t-1)\gamma_{ex}). \qquad (2.2)$$

As a viable alternative, Mazur and Hastie (1978) have proposed a hyperbolic function, the core function of which is $g_{hy}(t) = \frac{t-1}{-t+1-\gamma_{hy}^{-1}} + 1$. Combined with Equation (2.1), this gives

$$f_{hy}(t) = \alpha_{hy} - (\alpha_{hy} - \beta_{hy})\left(\frac{t-1}{-t+1-\gamma_{hy}^{-1}} + 1\right), \qquad (2.3)$$

As a third function describing learning, a power curve has been forwarded by, for example, Logan (1988, 1995) and, in its more general form, Newell and Rosenbloom (1981). The core function of the simple power curve is $g_{po}(t) = t^{-\gamma_{po}}$, while for the general power curve it is $g_{gpo}(t) = (t+\delta)^{-\gamma_{gpo}}$, where the additional parameter δ takes into account learning prior to the beginning of the task. If the core function of the simple power curve is substituted into Equation (2.1), we have

$$f_{po}(t) = \alpha_{po} - (\alpha_{po} - \beta_{po})t^{-\gamma_{po}}. \qquad (2.4)$$

The different core functions mentioned above do not only affect the curvature of learning trajectories, but they also have important theoretical implications. For example, as detailed by Restle and Greeno (1970), the exponential curve may be interpreted as being based on a "replacement model" of learning. It suggests that learning is a process through which incorrect response tendencies are replaced by more and more correct response tendencies. As an exponential curve, the replacement model implies a constant learning rate relative to the amount left to be learned, that is, the replacement process is assumed to occur at a constant rate. As such, the exponential learning model fits nicely into the theories of Estes (1950) and Bush and Mosteller (1955). By contrast, as Restle and Greeno (1970) pointed out, the hyperbolic curve is based on a "accumulation model" of learning. According to the accumulation model, learning is a process by which correct response tendencies increase steadily with practice and compete with incorrect response tendencies, which remain constant across trials. Unlike the exponential model, the amount of accumulation per trial is considered a constant proportion of the amount or duration of the study. The accumulation model was first introduced by L. L. Thurstone (1919) in his monograph on the learning curve. Finally, the power curve is based on the assumption that ". . .some mechanism is slowing down the rate of learning" (Newell & Rosenbloom, 1981, p.18). Thus, if learning follows a power law, learning slows down across trials. This slowing, however, is not proportional to the amount left to be learned or the duration of study. ACT-R (Anderson & Lebiere, 1998) and SOAR (Newell, 1990), two cognitive architectures, generally predict a power law of learning, albeit for different reasons. ACT-R posits that rules and memory traces are strengthened across trials according to a power law based on the assumption that the cognitive system is adapted to the statistical structure of the environment (Anderson & Schooler, 1991). Several models in SOAR have been created that model the power law (e.g., Nerb, Ritter, & Krems, 1999; Newell, 1990). These models explain the power law as arising out of mechanisms such as hierarchical learning (i.e., learning parts of the environment or internal goal structures) that initially comprises low-level actions being very common and, thus, useful. With further practice, even more valuable, larger patterns of actions that occur less frequently are learned.

The task of the present chapter is not, however, to decide which type of learning curve is the "true" one for verbal learning in old age. In the end, the issue of which core function

describes learning most adequately—be it in the elderly or other age groups—is still controversial (e.g., Heathcote et al., 2000; Logan, 1995; Mazur & Hastie, 1978; Newell & Rosenbloom, 1981; Newell, Liu, & Mayer-Kress, 2001).[9] Rather, we aimed at extending the examination of learning curves in old age by a thus far neglected dimension: While previously, learning curves have almost exclusively been investigated using averaged data (e.g., Logan, 1988), we wanted to bring the individual back into the investigation of learning curves. Specifically, we intended to model individual differences in the three parameters governing learning curves in old age as described in Equations (2.2, 2.3, and 2.4). In our opinion, it seems unwise to leave information regarding individuals unused in examining learning curves, a point that has similarly been made with respect to developmental changes in the elderly (cf. Hofer & Sliwinski, 2001, 2006; Zimprich, 2002; Zimprich & Martin, 2002).

Assessing the amount of verbal learning across a number of practice trials necessarily requires repeated measurements. In this respect, the examination of systematic performance changes due to learning bears similarities to investigating developmental changes over time. However, while the study of developmental changes has benefited from novel statistical analysis techniques that reach beyond the traditional analysis of variance approach, for example, growth curve models that distinguish between "fixed" or average effects and "random" or individual effects (cf. Bryk & Raudenbush, 1992, chap. 6; Goldstein, 1995, chap. 6; McArdle & Anderson, 1990), the same has not happened regarding learning curves. Notably, a key feature of these comparatively recent statistical approaches is that by including random effects they allow for modelling interindividual differences in intraindividual change and the inclusion of explanatory variables that may account for the diversity in longitudinal trajectories (e.g., Zimprich, 2002; Zimprich & Martin, 2002). By contrast, the analysis of age differences in learning curves is still dominated by statistical approaches relying mainly on means, that is, average performance changes, where individual differences in memory performance increments are treated as nuisance (cf. Davis et al., 2003).

In what follows, we try to demonstrate the fruitfulness of the individual-centered approach on verbal learning in old age outlined in Chapter 1.4.2.2.

[9] During the 1950s, different models of learning (Bush & Mosteller, 1955; Estes, 1950) oftentimes fit the data almost equally well. The reason for this was that empirical predictions derived from the various models were rather similar. A similar problem may be observed in distinguishing among the exponential, hyperbolic, and power functions using only limited amounts of data.

2.2.1 An empirical analysis of verbal learning in old age

The data used in the sequel come from the Zurich Longitudinal Study on Cognitive Aging (ZULU), an ongoing longitudinal study on cognitive and learning abilities of elderly persons in Switzerland (Zimprich et al., in revision). At first measurement occasion (T1: 2005), the sample of the Zurich Longitudinal Study on Cognitive Aging (ZULU) comprised 364 participants who where between 65 to 80 years of age (Mean age: 73 years, SD = 4.4 years; 46% women). The majority of the sample was married and resided with others. On average, participants reported about 13 years of formal education. For the sample, there were no signs of cognitive impairments or pronounced depressive affect. The majority of participants judged their health as ”good” and, in addition, no participant reported any severe hearing or vision difficulties. Part of the cognitive testing protocol in ZULU were three measures of processing speed (Number Comparison, Identical Pictures, Letter Digit Substitution), a verbal learning measure that comprised five trials of a word list recall task, and three measures of memory (Paired Associates, Story Recall, Picture Memory).

Number Comparison (Ekstrom, French, Harman, & Dermen, 1976) required participants to compare as rapidly as possible whether two numbers presented on the computer screen were identical or not. Scored was the number of correct answers, which could range between 0 and 60. After two practice items during the instruction phase, the time to work on the task was 90 seconds. Identical Pictures (Ekstrom et al., 1976) required participants to choose one out of five objects that was identical to a reference object as rapidly as possible. Scored was the number of correctly answered items, which could range from 0 to 60. After two practice items during the instruction phase, the time to work on the task was 90 seconds. Eventually, Letter Digit Substitution consisted of 75 items. For each item, a table that assigned five different letters to the numbers one to five was displayed on the top of the computer screen. Below the table, a single letter was presented together with a question mark. Participants were supposed to press the number that belonged to the single letter according to the presented coding table. For each item, there was a different coding table. Scored was the number of correctly answered items, which could range from 0 to 75. After two practice items, participants had 90 seconds to work on the task.

Verbal Learning was assessed by five consecutive trials of a word list recall task. The task comprised 27 meaningful, but unrelated words that were taken from the German Version of Rey Auditory Verbal Learning Test (Helmstädter, Lendt, & Lux, 2001). The 27 words

appeared on a computer screen at a rate of two seconds each and participants were required to read them aloud. After the presentation of all 27 words, participants were asked to recall as many words as possible in any order. This procedure was repeated five times, with the order of words being different for each trial. At each trial, the number of correctly recalled words was scored, ranging between 0 and 27.

Paired Associates comprised 12 semantically unrelated word pairs taken from the German version of the Wechsler Memory Scale-Revised (WMS-R: Härting et al., 2000) and from the Munich Verbal Memory Test (MVGT: Ilmberger, 1988). After two examples during instruction, word pairs were presented for four seconds each and participants had to read them aloud. Following a pause of one second, the next word pair was displayed. After presentation of all 12 word pairs, only the first word of a pair appeared on the screen as a cue and the second one was replaced by a question mark (e.g. salad - ?), using a different order than during encoding. Scored was the number of correctly recalled target words, which could range from 0 to 12. Story Recall consisted of story A of the Logical Memory subtest of the German version of the Wechsler Memory Scale-Revised (Härting et al., 2000). The 66-word story was read by the experimenter during 60 seconds. Participants were asked to listen closely and, when the story was finished, to immediately recall as many details as possible. Scored was the number of correctly recalled propositions, which could range from 0 to 25. Finally, Picture Memory encompassed 12 pictures taken from the Nuremberg Age Inventory (Nürnberger-Alters-Inventar: Oswald & Fleischmann, 1999). For each item, a picture of a simple object for 2.75 seconds and participants were required to name the shown object aloud (e.g., “apple”). Followed by a pause of one second, the next picture was displayed. Immediately after presentation of all 12 pictures, participants were asked to verbally recall as many of the seen objects as possible. Scored was the number of correctly recalled objects, which could range between 0 and 12.

All analyses reported below were conducted using *Mx* (Neale, Boker, Xie, & Maes, 2003). Nonlinear learning models were specified as structured growth models (Browne, 1993; Browne & Du Toit, 1991). The absolute goodness-of-fit of models was evaluated using the χ^2-test and two additional criteria, the Comparative Fit Index (CFI) and the Root Mean Square Error of Approximation (RMSEA). Values of the CFI above .95 are considered to be adequate, whereas for the RMSEA values less than .06 indicate an acceptable model fit (cf. Hu & Bentler, 1999). In comparing the relative fit of nested models, we used the χ^2-difference

test where appropriate. Due to its dependency on sample size and due to the fact that we also wanted to compare non-nested models, we mainly relied on calculating 90% RMSEA confidence intervals for the models estimated (MacCallum et al., 1996). Because the RMSEA is virtually independent of sample size, the comparison of RMSEA confidence intervals, that is, whether they do or do not overlap, provides an effective, alternative method of assessing relative model fit of nested and non-nested models. Throughout, we refer to a significance level of $p < .05$ if a parameter estimate is denoted as statistically significant.

2.2.2 Emprirical findigs

Descriptive statistics of the 11 manifest cognitive variables and age together with their intercorrelations are shown in Table 2.5. For the means of the verbal learning indicators, a typical learning curve emerged, that is, a gradually increasing, but negatively accelerated trajectory. Raw data were checked for departures from both univariate and multivariate normality. Skewness and kurtosis estimates of the 11 manifest cognitive variables did not exceed 1 or −1 (average skewness 0.08; average kurtosis 0.25), whereas the distribution of age, for which the sample had been stratified, was platykurtic. The normalized estimate of Mardia's coefficient of multivariate kurtosis was 0.65. Thus, with the limitation that the distribution of age was inconsistent with univariate normality, the multivariate distributional properties of the 11 manifest cognitive variables and age warranted the use of maximum likelihood parameter estimation.

Models of Verbal Learning in Old Age

In a first model (VL1), changes in verbal learning performance were fitted by a growth curve comprising level, linear slope, and quadratic slope. As can be seen from Table 2.5, Model VL1 evinced a satisfactory fit as indexed by the CFI, although not so as judged by the statistically significant χ^2-value and the RMSEA. On average, 82% of variance were explained in the verbal learning indicators. For the latent level variable, a mean of 5.34 was estimated, while for linear slope and quadratic slope they were 4.43 and −.45, respectively. Variances were estimated as 3.93 (level), 4.46 (linear slope), and 0.06 (quadratic slope). These variances were all statistically significant, implying that there were reliable individual differences in initial performance level, linear change across trials, and negative acceleration of performance changes across trials.

Table 2.5: *Descriptive statistics and sample correlations of cognitive variables and age*

	Mean	Std	1	2	3	4	5	6	7	8	9	10	11
1. Number Comparison (NC)	17.15	4.3											
2. Identical Pictures (IP)	21.38	4.54	0.49										
3. Letter Digit Substitution (LD)	31.86	6.62	0.64	0.63									
4. Verbal Learning, Trial 1 (L1)	5.27	2.24	0.18	0.24	0.29								
5. Verbal Learning, Trial 2 (L2)	9.62	3.15	0.28	0.31	0.36	0.68							
6. Verbal Learning, Trial 3 (L3)	12.33	3.15	0.31	0.29	0.41	0.62	0.81						
7. Verbal Learning, Trial 4 (L4)	14.38	4.03	0.29	0.29	0.41	0.57	0.77	0.85					
8. Verbal Learning, Trial 5 (L5)	16.01	4.47	0.23	0.25	0.35	0.53	0.73	0.81	0.85				
9. Paired Associates (PA)	2.99	2.35	0.14	0.15	0.21	0.23	0.38	0.46	0.45	0.44			
10. Story Recall (SR)	14.52	4.12	0.15	0.15	0.25	0.21	0.28	0.33	0.33	0.28	0.33		
11. Picture Memory (PM)	6.84	1.57	0.23	0.29	0.34	0.33	0.44	0.46	0.45	0.43	0.23	0.23	
12. Age	72.98	4.43	- .25	- .35	- .37	- .12	- .16	- .20	- .17	- .12	- .18	- .07	- .21

Note. N = 364. Correlations larger than r = .09 in absolute size are statistically significant at $p < .05$ (one-tailed).

Level and linear slope were significantly related (r = .29), as were level and quadratic slope (r = −.30), and linear and quadratic slope (r = −.91).[10] Thus, those starting out at a higher level had a somewhat higher linear performance increase and a more pronounced negative acceleration in performance changes. Also, those with a high linear performance increase showed a much stronger flatting out of performance improvements. In sum, the quadratic growth model appeared to capture important aspects of verbal learning, but did not fit acceptably.

In a second model (VL2), exponential learning curves as described by Equation (2.2) in the introductory section were imposed. Model VL2 had an acceptable fit (see Table 2.5), with the χ^2-value indicating no statistically significant differences between the moments predicted by Model VL3 and actual moments of the data. Albeit a χ^2-difference comparison of Model VL2 with Model VL1 is impossible due to both models having the same degrees of freedom, the CFI as well as the RMSEA clearly favored Model VL2. Note, however, that the RMSEA confidence intervals overlapped somewhat, indicating that difference in fit was not statistically significant. The amount of explained variance in the verbal learning indicators was 84%, on average. The latent variable representing initial performance level (β_{ex}) had a mean of 5.29, while the latent variable reflecting potential maximum or asymptotic performance (α_{ex}) was 19.23, on average. Mean rate of learning (γ_{ex}) was estimated as 0.358. The statistically significant variances (with standard errors in parentheses) were 4.11 (0.69) for initial performance, 29.35 (7.60) for potential maximum or asymptotic performance, and 0.029 (0.012) for rate of learning. The correlation between initial performance level and potential maximum performance reached statistical significance (r = 0.43), implying that those starting out at a higher level of memory performance also tended to show a higher asymptotic memory performance after five learning trials. By contrast, the associations between initial performance level and rate of learning (r = −.22), and between rate of learning and potential maximum performance (r = .15) were statistically not significant. Taken together, the exponential learning curve model seemed to adequately describe the verbal learning data.

[10] Note that the strong correlation between linear and quadratic slope represents a statistical necessity and could be reduced by using, for example, orthogonal polynomial contrasts. Biesanz et al. (2004), however, have cautioned against doing so, because interpretation of parameters then may become meaningless.

Table 2.5: *Sequence of Estimated Models and Fit Statistics*

Model[a]	χ^2	df	*p*	CFI	RMSEA	90% CI
VL1 (Quadratic Learning)	30.90	6	< .05	.984	.107	.071–.145
VL2 (Exponential Learning)	8.37	6	> .20	.999	.033	.000–.081
VL3 (Hyperbolic Learning)	3.28	6	> .77	1.000	.000	.000–.046
AVL1 (VL3 & Age)	5.64	8	> .69	1.000	.000	.000–.048
AVL2 (VL3 & equal Age)	9.23	10	> .51	1.000	.000	.000–.054
SVL1 (AVL1 & Speed)	21.73	22	> .47	1.000	.000	.000–.043
SVL2 (SVL1 & equal Speed)	28.88	27	> .36	.999	.014	.000–.044
SVLM1 (SVL1 & Memory)	56.26	47	> .17	.996	.023	.000–.043
SVLM2 (SVLM1 - direct eff.)	62.05	49	> .10	.994	.027	.000–.046

Note. $N = 364$. df = degrees of freedom, CFI = Comparative Fit Index, RMSEA = Root Mean Square Error of Approximation, CI = Confidence Interval.

[a] See text for a more detailed description of the estimated models.

As an alternative to exponential verbal learning, we also fitted a hyperbolic model (VL3) of verbal learning as described by Equation (3) in the introductory section. The fit of Model VL3 was excellent (see Table 2.5). Again, a χ^2-difference comparison with the previous model (VL2) was impossible, but both the CFI and RMSEA showed a better fit of Model VL3. At the same time, the overlapping RMSEA confidence intervals indicated that the difference in model fit was not statistically significant. However, RMSEA confidence intervals in comparison with Model VL1 did not overlap, denoting that Model VL3 fitted significantly better than Model VL1. On average, 84% of variance were explained in the five manifest indicators of verbal learning. For the latent variable capturing initial performance level (β_{hy}), a mean of 5.28 emerged. The latent variable representing potential maximum performance (α_{hy}) had a mean of 26.58. Note that this estimate was much closer to the actual list length (27), that is, the maximum number of words that could be recalled, than in the exponential model (19.23). The mean rate of learning (γ_{hy}) was 0.251. Variances (with standard errors in parentheses) of the learning parameters were 4.23 (0.87) for initial level, 84.83 (28.78) for potential maximum performance, and 0.026 (0.012) for the rate of learning. Hence, with respect to all three parameters there were reliable interindividual differences. The

nonsignificant correlations of initial performance with potential maximum performance and with rate of learning were r = .25 and r = .10, respectively. The statistically significant correlation between potential maximum performance and rate of learning was r = −.65, indicating that those with a higher asymptotic performance had a slower rate of learning, that is, they needed more trials to achieve their potential maximum performance. In sum, the hyperbolic model of learning captured the data very well.

Eventually, a power model of learning as given by Equation (2.4) was estimated. However, we were unable to arrive at a solution that led to a stable estimate of potential maximum performance (α_{po}), which was estimated as being 133. As a consequence of these estimation difficulties, a number of parameters became statistically nonsignificant. More specifically, upon inspection, the power curve fitted excellently within the range of the five trials providing data, but afterwards hardly changed its slope, which led to the although formally correct, but unstable estimate of α_{po} and its variance. Probably, with some more trials, we would have been able to estimate α_{po} consistently. Based on these difficulties in estimation, however, we decided to skip the power model of learning from further analyses.

To summarize, it appeared that a hyperbolic model represented the ZULU verbal learning data best, because it showed the best point estimates of model fit. In addition, the fit of the hyperbolic model was significantly better than that of the quadratic model. Hence, we decided to accept and retain the hyperbolic model for the analyses to follow. Note, however, that due to the very similar form of the trajectories in the first five trials, we can not safely conclude that the hyperbolic model outperforms the exponential model. Thus, using the ZULU data, it was impossible to distinguish between the replacement and accumulation models of verbal learning, because neither one of them does clearly outperform the other in terms of model fit.

Figure 2.2 depicts the predicted trajectories as based on Model VL3. Shown are seven randomly selected model-based trajectories (thin lines) and the mean trajectory (thick line), dots denote observed values, circles denote observed means. As can be seen, Model VL3 does very well in describing, of course, the mean learning curve, but also in capturing individual learning curves.

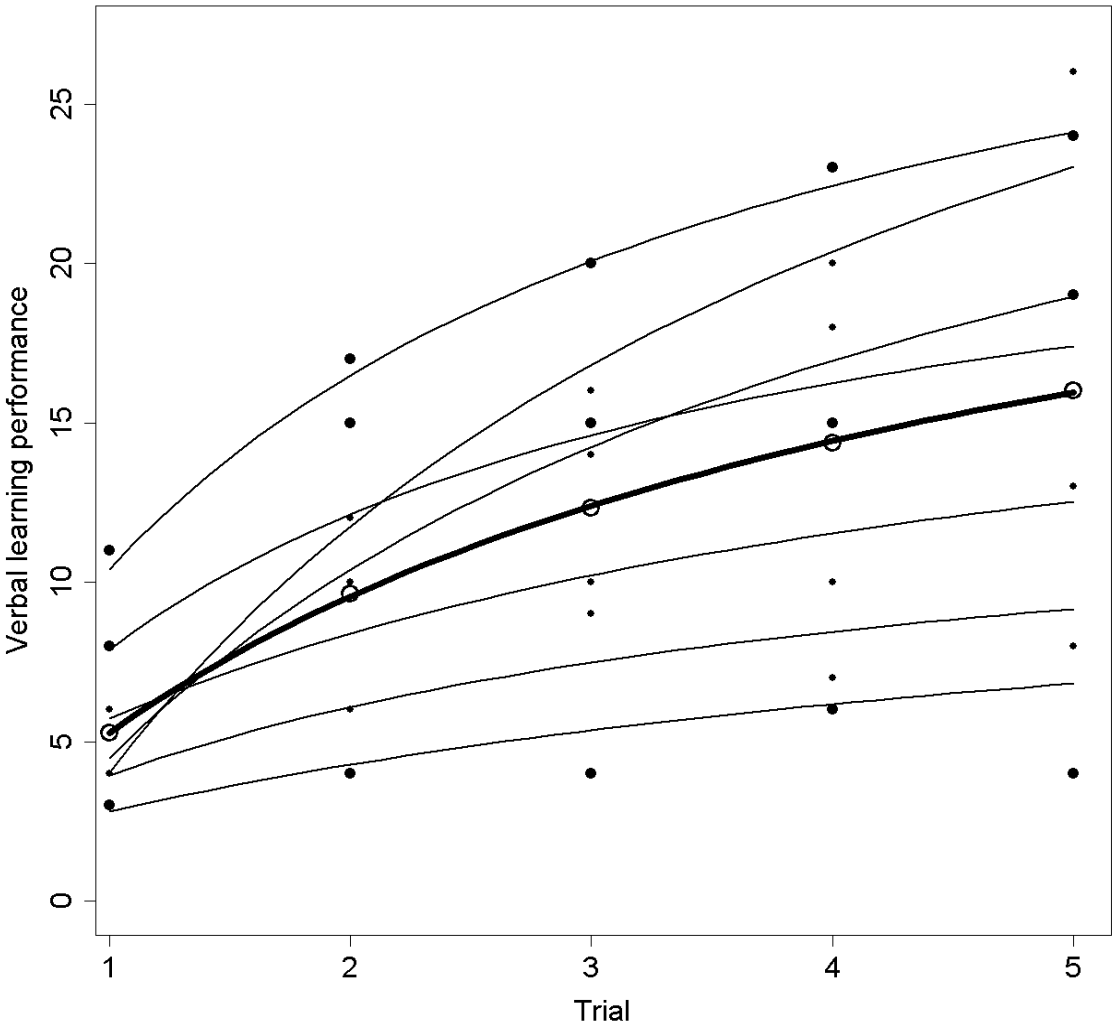

Figure 2.2: Shown are seven randomly selected model-based trajectories. The thick black line denotes the mean profile. Predicted values are based on the hyperbolic model (VL3), dots represent observed values.

Covariates of Learning in Old Age

As a first extension of the hyperbolic learning model, age was included as a predictor of individual differences in the three learning parameters initial level (β_{hy}), potential maximum performance (α_{hy}), and rate of learning (γ_{hy}). This extended model (AVL1) achieved an excellent model fit (see Table 2.5), which was virtually identical to that of Model VL3, implying that age effects on the manifest indicator variables of verbal learning were mediated completely by the three learning parameters. The standardized effect of age on β_{hy} was −0.13 and statistically significant, accounting for approximately 2% of variance in initial

level—a small effect in terms of the standards recommended by Cohen (1988). By contrast, the standardized regression of α_{hy} on age was 0.07, a value so small that it did not reach statistical significance. Accordingly, age explained approximately 0.5% of variance in potential maximum performance. The significant standardized age effect in γ_{hy} was estimated as −0.23, which amounted to 5% of explained variance (or an effect of medium size) in rate of learning. Thus, it appeared that age accounted mainly for individual differences in rate of learning, followed by the effect on initial performance level. Effects were in the medium to small range, however, indicating that individual differences among participants of the same age by far outweighed the age-related differences.

In an attempt to more rigorously test for differences in the regression of the three learning parameters on age, in the next model (ALV2) the standardized regression coefficients were constrained to be equal. As shown in Table 2.5, imposing these constraints did hardly reduce model fit. Albeit the point estimates in Model AVL1 indicated different age-related effects, these differences were not reliable. The constrained standardized regression parameter was −0.08 and statistically significant. Based on the ZULU data one may not safely conclude that age had a differential effect on initial level, potential maximum performance, and rate of learning. An explanation for this lack of statistical power is that differences in age-related effects were small, which, after taking into account the sample size of 364, made it virtually impossible to differentiate Model AVL2 from AVL1—in fact, as calculated using the procedure suggested by McCallum, Browne, and Sugawara (1996), power was 0.02 only.

Next, for Model SVL1, processing speed was included as an additional predictor of the three learning parameters into the AVL1 model. Processing Speed was measured by three manifest variables, namely Number Comparison, Identical Pictures, and Letter Digit Substitution. Standardized factor loadings on the common speed factor were .69 (Number Comparison), .70 (Identical Pictures), and .91 (Letter Digit Substitution), indicating large amounts of shared variance of the three indicator variables of speed of information processing. The standardized regression of processing speed on age was −.42, with the latter accounting for 17% of variance in the former. As displayed in Table 2.5, Model SVL1 fitted the data excellently. Processing speed showed statistically significant effects on initial performance level (β_{hy}), followed by learning rate (γ_{hy}), with standardized regression coefficients of .34 and .25, respectively. That is, those who processed information more rapidly remembered more words at the beginning of the verbal learning test and showed a

more pronounced learning rate compared to individuals with low processing speed. By contrast, the effect of processing speed on potential maximum performance (α_{hy}) was not significant (.13). At the same time, the effects of age on the three learning parameters were attenuated to statistical non-significance. Specifically, processing speed mediated 87% of the age-related effects in initial performance, 42% in potential maximum performance, and 43% in rate of learning. The correlation between potential maximum and learning rate increased slightly to $r = -.66$ and remained statistically significant. The amount of variance processing speed and age accounted for ranged, in terms of effect size, from small to medium, with 11% in β_{hy}, 2% in α_{hy}, and 10% in γ_{hy}. Together, processing speed and age thus exerted medium effects on initial performance level and learning rate, whereas potential maximum performance seemed largely unaffected.

As with the effects of age (see Model AVL2), one may wonder whether the effects of processing speed on β_{hy}, α_{hy}, and γ_{hy} were significantly different. Hence, after removing the non-significant direct effects of age on initial level, potential maximum, and learning rate, we constrained the standardized effects of processing speed on the three learning parameters to be equal in Model SVL2. Table 2.5 shows that doing so did not lead to a substantial decrement in model fit. The constrained regression parameter was .31 and statistically significant. Processing speed explained approximately 9% of variance in initial performance level, potential maximum performance, and rate of learning. As a consequence, one should not consider the effects of processing speed on β_{hy}, α_{hy}, and γ_{hy} as being statistically different in the population. However, taking into account the small differences in model fit and the ZULU sample size, it was practically impossible to differentiate Model SVL2 from SVL1: Statistical power was only 0.05.

The next step was to include memory as an outcome variable of verbal learning (SVLM1). Memory was assessed by three manifest variables: Paired Associates, Story Recall, and Picture Memory. The standardized factor loadings on the memory factor were .54 (Paired Associates), .42 (Story Recall), and .54 (Picture Memory), showing that memory indicators shared substantial amounts of variance, albeit less than the processing speed indicators did. Model SVLM1 evinced an excellent model fit (see Table 2.5). The strongest effect on memory was exerted by the learning rate, with a standardized regression coefficient of .84, followed closely by potential maximum performance (.83) and initial performance level (.21). Hence, those with a higher rate of learning showed better memory performance, which was

also true for those with a higher potential maximum of verbal learning and, albeit to a much lesser extent, for those high in initial performance level. Note that memory was assessed by three one-trial tests, from which one might have expected that initial verbal learning performance (β), which, in terms of its definition, is most similar to memory, should emerge as the strongest predictor. However, initial performance level turned out to be the least predictive learning parameter. The standardized effect of age on memory (−.11) did not reach statistical significance, neither did the effect of processing speed (.13). Note that Model SVLM1 can be regarded as a mediational model, in which verbal learning mediated the effects of processing speed on memory. More specifically, the three learning parameters mediated approximately 70% of the effect of processing speed on memory. Together, age, processing speed, and the three learning parameters β_{hy}, α_{hy}, and γ_{hy} accounted for sizable 85% of variance in memory.

Next, for Model SVLM2, the non-significant direct effects of age and processing speed on memory were removed. Table 2.5 shows that Model SVLM2 achieved an excellent fit, which, compared to Model SVLM1, was not statistically significant ($\Delta\chi^2$= 5.79, df = 2, p > .05). At the same time, the effects of processing speed on β_{hy}, α_{hy}, and γ_{hy} changed somewhat in proceeding from Model SVLM1 to SVLM2: While the explained variance in the three learning parameters remained virtually unchanged in initial level and potential maximum performance, it increased to 28% in the learning rate. That is, the three learning parameters acting as mediating variables affected also the association strength between processing speed and learning—a result that might appear counterintuitive, but represents as statistical necessity (cf. Pedhazur, 1982). The correlation between potential maximum and initial performance reached statistical significance (r = .29), whereas the correlation between potential maximum performance and learning rate was not significant any longer and decreased to r = −.43. In Model SVLM2, the standardized effect β on memory was unaltered, while the standardized effect of α decreased somewhat (.63) and that of the learning rate increased slightly (.85). The amount of explained variance in memory increased to 92%.

We selected Model SVLM2 as representing the interrelations among age, processing speed, memory, and the three parameters of verbal learning adequately, while being as parsimonious as possible. Model SVLM2 is depicted in Figure 2.3.

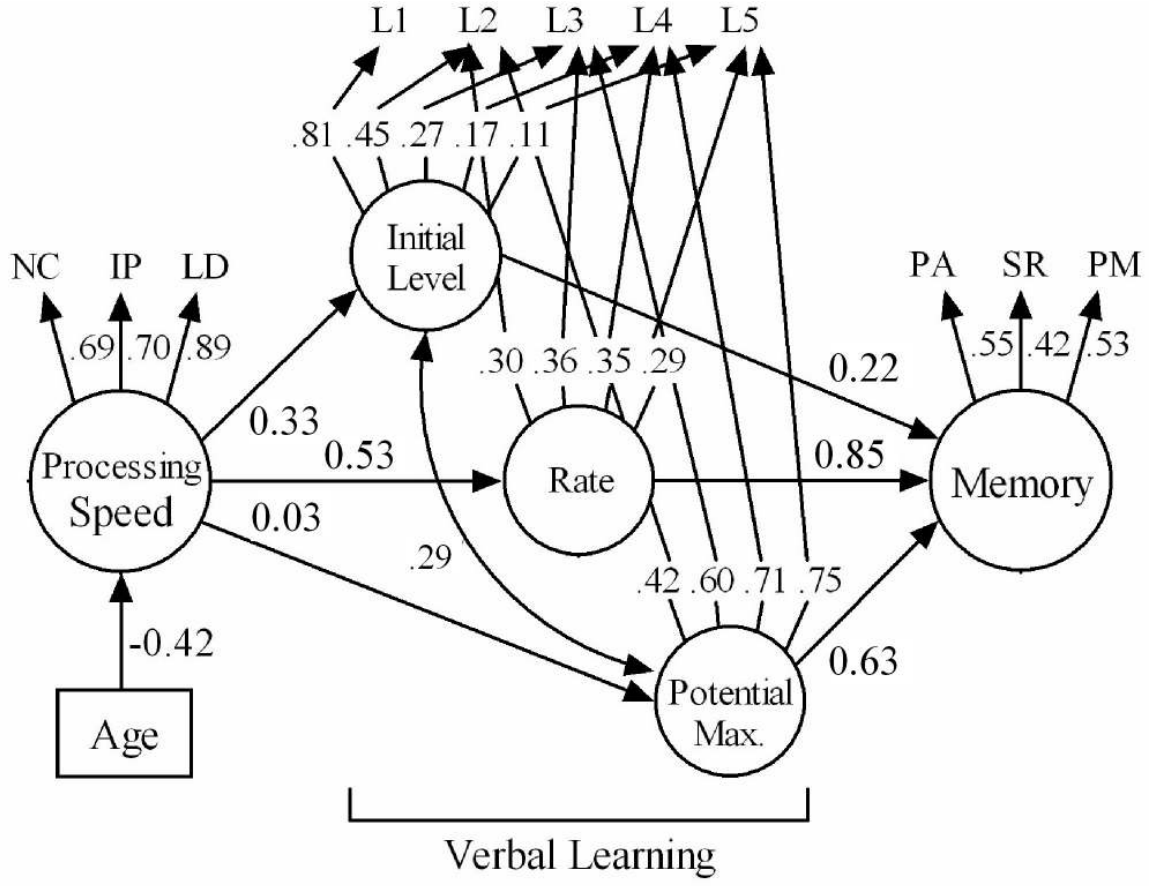

Figure 2.3: *N* = 364. All parameters are standardized. NC = Number Comparison, IP = Identical Pictures, LD = Letter Digit Substitution, L1–L5 = Verbal Learning Indicators, PA = Paired Associates, SR = Story Recall, PM = Picture Memory.

2.2.3 Conclusions

For the present investigation, we set out to bring the individual back into the verbal learning curve, an issue that has been neglected from our point of view. Focusing on the average learning curve only and, thus, relegating individual differences in verbal learning to a nuisance parameter appears to be antithetical to a science of development. Before we turn to the substantive implications of this individual-focused approach, a short discussion of methods for capturing individual learning curves seems in order.

The fact that the average learning curve can be different and can even have a different functional form than the majority of individual curves represents a well-known result (e.g., Sidman, 1952). The situation becomes more complicated yet if one acknowledges that learning performance can be averaged across persons, across blocks of trials, or both, and that

each way of summarizing data gives rise to specific difficulties (Brown & Heathcote, 2003; Cousineau, Hélie, & Lefebvre, 2003).Hence, a curve that describes grouped data must not necessarily be representative of any individual person. One way to bring back the individual into learning curves is, thus, to focus on individual data, that is, model learning directly for single persons. Heathcote and colleagues (2000), for example, did so in investigating whether the exponential or the power function is more appropriate in describing learning data from a variety of studies. This approach, however, holds the shortcoming that standard errors of parameters may be biased. Also, in estimating individual parameters it does not use information provided by other individuals, whose trajectory is similar. A natural way to strike a balance between grouped data and individual data is to borrow strength from both sides. For this reason, we chose a structured growth curve approach. In this sense, our perspective on learning curves resembles nonlinear mixed effects models as described in Davidian and Giltinan (1995) or Molenberghs and Verbeke (2005, chap. 20), although for actually fitting curves to data we used structured latent curve models as developed by Browne (1993; Browne & Du Toit, 1991), which allows for fitting nonlinear growth curves belonging to the Richards (1959) family as structural equation models. From the perspective of developmental research in old age, extending the investigation of learning curves in old age to models incorporating individual effects is only natural. Developmental aging research has become aware of individuals again some years ago, based on the pioneering work of methodologists who, in the late 80s and early 90s, provided developmentalists with the tools needed to model individual differences (Bryk & Raudenbush, 1992; Goldstein, 1995; McArdle & Anderson, 1990). This has resulted in new insights into the process of aging and, at the same time, raised new questions about development that require theoretical elaboration.

After having clarified these methodological points, we turn to the substantive issues regarding verbal learning in old age. What have we learned about verbal learning in old age by bringing back the individual into the learning curve? Among the four functions applied, the hyperbolic learning curve seemed to describe the evolvement of performance across the five trials best. Each of the three parameters of the hyperbolic curve can be interpreted in a different way: While β_{yh} denotes the initial level in learning performance, α_{hy} is a more theoretical value since it is formulated as the upper asymptote or limiting value. Hence, it can only be approached, but never achieved within a given range of trials. Finally, γ_{hy} defines the curvature of the learning trajectory, that is, a high learning rate leads to a steep increase in

learning performance across the first trials, whereas a low rate leads to a flatter trajectory and a more evenly distributed increase in learning (Browne, 1993; Browne & Du Toit, 1991; Meredith & Tisak, 1990). The hyperbolic function would, strictly speaking, also mean that the amount of accumulation per trial is a constant proportion relative to the trials completed (Mazur & Hastie, 1978). In the ZULU sample, however, fit of the hyperbolic and the exponential function was statistically indistinguishable, which leaves the question of whether learning follows an accumulative or a replacement process an open issue.

Note that the excellent model fit implied that *both* the average learning curve and the individual ones were captured by the hyperbolic equation—albeit with varying parameter values. That is, in the verbal learning data we analyzed, the average curve and individual curves were of the same functional form. Fitting the hyperbolic model, thus, allowed confirming that verbal learning showed reliable interindividual differences in old age. In order to compare the magnitude of individual differences in the parameters β_{hy}, α_{hy}, and γ_{hy}, we calculated the coefficient of variation (CV) of each parameter. These amounted to CV(β_{hy}) = 0.39, CV(α_{hy}) = 0.35, and CV(γ_{hy}) = 0.64, implying that individual differences were, relatively seen, most pronounced in rate of learning. Hence, older people tended to show more pronounced individual differences from each other in the rate of acquisition than in the initial level or total potential maximum performance. The random effects potential maximum performance (α_{hy}), and rate of learning (γ_{hy}) were negatively correlated, implying that those with a higher upper limit of learning performance needed more trials to bridge the performance gap between initial performance level and maximum performance. We acknowledge, though, that this latter finding might also be indicative of a ceiling effect, although our list comprised 27 words, which, together with the fact that the mean number of recalled words was 16 at Trial 5 (see Table 2.5), renders strong ceiling effects unlikely. Still, more detailed analyses are required in this regard, for example, by excluding participants with a performance close to the maximum number of words.

Next, we examined the relations between aging and the three verbal learning parameters. The finding that aging negatively affected verbal learning performance as a whole is not new (see Kausler, 1994), but due to the latent growth curve approach, we were able do refine Kausler's observations in several respects. According to the point estimates of effects, aging mainly affects the verbal learning rate, followed by the effect on initial performance, while potential maximum performance was practically unrelated to age. Thus, with increasing

age, more learning experiences were needed to attain the same level of mastery. These findings would imply that age differences in verbal learning are due to a flattening of the learning curve as one grows older, which also has the effect that the (predicted) learning curves of some individuals look almost linear (see Figure 2.2). One has to take into account, however, that the ZULU sample consists of elderly persons only and that age-related effects were in the small-to-medium range. It remains an open issue, at present, whether larger age-related effects in verbal learning parameters would have been found using a broader age range. We certainly would expect this to happen for initial performance level, because this parameter closely mirrors the typical one-trial memory assessments shown to follow a decline trajectory in old age (Craik, 1977; Davis et al., 2003; Hultsch et al., 1998; Kausler, 1994). But, at current, we do not know how much variance age might account for in α and γ in a lifespan sample. A more age-heterogeneous sample—and, thus, potentially stronger age-related effects—might also help overcoming the lack of statistical power we faced in our analyses when age effects were constrained to be equal (cf. MacCallum, Browne, & Sugawara, 1996). Thus, we think that it is important to follow this research path further, because it might confirm what corresponds to a common lay impression of aging: Older persons take longer to learn, but can, given enough effort and time, reach the same level of mastery as younger adults do. Having said this, the small age-related effects in verbal learning necessarily mean that, in the ZULU data, individual differences among persons of the same age by far outweighed age-related differences in β, α and γ. As an alternative research avenue—and more in line with the focus of the present chapter—, we would thus like to encourage researchers to focus on individual differences orthogonal to cross-sectional age (cf. Zimprich et al., 2007). In line with this, another fruitful approach would be to analyze the development of the three parameters longitudinally (Hofer & Sliwinski, 2001, 2006).

Subsequently, processing speed was included as an additional predictor of verbal learning. Based on the assumption that speed of information processing is more basic than other cognitive abilities and, hence, represents a resource for higher order cognitive functions (Salthouse, 1991), it appeared instructive to examine which of the three learning parameters was most strongly affected by speed. It turned out that processing speed had positive, medium-sized effects on initial performance level and learning rate, while the effect on potential maximum performance was small. Thus, those elderly persons higher in mental speed started out at a higher level and, more importantly, showed a steeper increase in their

verbal learning trajectories. These findings underline the importance of being able to process information rapidly for remembering new material after one trial—a finding that is well-established in the literature on memory aging (e.g., Hultsch et al., 1998). In addition, processing speed appeared even more important for the number of trials needed to attain one's potential maximum learning performance, that is, rate of learning. Although a more stringent test on the equality of processing speed effects on learning parameters showed that they were statistically indistinguishable, this might also be the consequence of our medium-sized sample. Hence, we believe that in a larger sample these effects may prove to be distinct. After including speed, the age effects on learning were no longer statistically significant, implying that the age effects in verbal learning were mediated by speed, which is in line with Salthouse's (1996) processing speed theory. However, one has to keep in mind that, as Hofer and Sliwinski (2001) have argued, tests of mediational hypotheses in models of cognitive aging might be problematic if they rely on cross-sectional (between-person variance) methods instead of longitudinal (within-person variance) methods, because the provide only a weak basis for drawing conclusions about correlated within-person age changes (see also Cole & Maxwell, 2003).

An important result we gained from the learning parameters was the finding that learning rate, followed by potential maximum performance, had the strongest effect on memory. In terms of its definition, one would have expected β to be the strongest predictor of memory because both are defined as the recall performance after one learning trial. Instead, initial learning displayed the smallest effect. This finding is even more intriguing if one considers the task proximity of the indicators for memory, especially Picture Memory or Paired Associates, which are procedurally and conceptually very similar to the verbal learning task (cf. Kausler, 1994). Unexpectedly, then, our results suggest that learning rate and potential maximum performance can be regarded as memory-inherent and highly predictive for memory performance. Regarding learning rate, a possible explanation might be that it is highly relevant for initial performance level as well, because learning, of course, already takes place before the first recall trial. Other parameterizations of the learning curve, which do not include a parameter for initial level, describe this situation of learning (with individually differing rates) right from the start more adequately than the ones we applied (see Mazur & Hastie, 1978). With respect to potential maximum performance, one might argue that verbal learning and testing-the-limits (see, e.g., Lindenberger & Baltes, 1995) share some

commonalities: Across trials, participants get closer to their specific performance limits, which increases individual differences. At the same time, these increasing individual differences can be mapped more exactly, because the full range of the measurement scale of 27 words is better utilized from trial to trial. Together with the assumption that closer to the limit, chance or unsystematic influences on performance become smaller, the reliability of measuring interindividual differences should increase, which should lead to stronger correlations with other variables, for example, memory. We also demonstrated that verbal learning mediated the direct effect of processing speed on memory to a considerable extent (70%). This finding might, at first glance, appear surprising, but in light of the fact that verbal learning almost completely accounted for the variance in memory, this strong mediational effect represents nearly a necessity.

Future Perspectives

We think we have demonstrated the usefulness of an individual-centered analysis of verbal learning in the preceding sections. The analysis of short-term repeated measures data by means of the hyperbolic equation including random effects we presented, offers researchers new possibilities to examine seemingly well established relations between cognitive constructs as, for example, verbal learning, speed, and memory (cf. Cudeck & Harring, 2007). We would argue that individual differences in verbal learning parameters as provided by the hyperbolic function are psychologically meaningful in that they capture between-person variability in within-person performance changes occurring at different stages of learning. Moreover, all three learning parameters exhibited substantial individual differences, implying that individual learning trajectories should not be collapsed across individuals because this would discount both theoretically and practically relevant information. Of course, our understanding of learning in old age would also benefit from transferring the analyses presented herein to other types of material and other types of learning, for example, skill acquisition (Ackerman, 1988; Cerella, Onyper, & Hoyer, 2006; Wisher et al., 1995).

The individual differences in verbal learning, however, require more elaborated conceptual models to explain and predict individual learning trajectories. We acknowledge that it is important to ask how and why people learn in old age (cf. Estes, 1950; Bush & Mosteller, 1955; Kausler, 1994), but would suggest complementing this question by asking of how and why people learn *differentially* in old age. Although we would expect that theoretical

approaches aiming to answer each question show a large overlap, they would still focus on different aspects. An advantage of a differential perspective on learning is that a number of explanatory variables, be it from the cognitive or from other domains, can easily be included as individual-differences variables. This also allows for a shift from ANOVA-type models to regression-type models. For example, one might expect that older persons high in typical intellectual engagement show superior learning compared to older adults being intellectually less active (Dellenbach & Zimprich, in press). Also, the investigation of learning bears the potential to bridge the gap between objective cognitive performance and subjective judgements of one's cognitive performance since learning may constitute a more naturalistic measure of memory (Rast et al., submitted; Zimprich et al., 2003).

The investigation of individual differences in learning is attractive also from a conceptual perspective of development: One might speculate that learning represents "microdevelopment," that is, development within a short time frame, as opposed to "macrodevelopment," which typically covers development over longer time spans. Lindenberger and Baltes (1995) conjectured that the mechanisms underlying learning might be the same or very similar to those underlying cognitive development, thus turning the study of learning into a showcase of examining cognitive development. Although development and learning are often treated as a dichotomy, they are both characterized by a persistent change of behavior over time, albeit time scales are different (cf. Newell et al., 2001). In accordance with such a link between learning and development, Zimprich, Hofer, and Aartsen (2004) have demonstrated that, in old age, learning at first measurement occasion is positively associated with the amount of longitudinal change in cognitive functioning. Integrating the examination of learning and development, thus, would bring together two research avenues that, once their different time horizons are taken into account, may turn out to be very similar.

In the same context, the analysis of learning curves as presented in the present chapter may be useful in describing retest effects in longitudinal studies (cf. Hofer & Sliwinski, 2006). To date, retest effects have oftentimes been taken into account in the form of comparatively unrealistic models, for example, by assuming that learning due to retest is linear or constant (e.g., Lövdén, Ghisletta, & Lindenberger, 2004). However, in order to disentangle two superimposed change processes in old age—one developmental process resulting in decline and one learning process leading to performance improvements—, one would either need specialized designs with, for example, independent samples, which

constitutes what has been done hitherto regarding retest effects. Alternatively, strong, testable hypotheses about the nature and form of the two processes at hand, development and learning, could be examined. To us, the latter approach now seems much more traceable: On the basis of the methods and analyses presented herein, it appears possible if not timely to revisit retest effects in longitudinal studies with a much stronger emphasis on learning than previously. Learning is not something to get rid off in longitudinal studies, but rather contains vital information about the cognitive aging process that awaits being utilized.

In closing, if one considers the verbal learning and memory duality a pendulum, we would encourage researchers to give this pendulum a new momentum such that it swings back into the direction of verbal learning. In the end, remembering new material is a dynamic process that, oftentimes, involves more than just one static learning cycle (cf. Nelson & Narens, 1994). In a broader sense, the present study illustrated the capabilities of nonlinear growth curve models as an analytical framework for linking both theoretical and methodological considerations in examining verbal learning and memory.

2.3 Age differences in the underconfidence-with-practice effect[11]

2.3.1 Introduction

A pertinent issue of research on metamemory is the degree to which individuals can accurately predict their memory performance. An accurate appraisal of one's own memorizing abilities seems desirable because, in subject to such an appraisal, more or less effort could be allocated to attain a certain level of mastery (Koriat et al., 2002; Nelson & Dunlosky, 1991). That is, based on their metacognitive judgments, individuals might be able to use self-monitoring to more efficiently control and regulate their strategies for learning and retrieving information from memory (Schneider & Pressley, 1989). The importance of monitoring may gain even more weight when cognitive resources and memory performance is declining, as it is the case with older adults, because this requires an optimized allocation of memory resources. An accurate appraisal of one's own memory functioning may essentially facilitate the appropriate allocation of cognitive resources in order to achieve a desired level of mastery and to avoid unnecessary overlearning. Findings regarding the changes in memory and metacognitive monitoring across the lifespan evidence that even though memory performance is impaired by aging, monitoring appears to be spared from cognitive decline (Connor et al., 1997; Hertzog et al., 2002; Shaw & Craik, 1989).

With respect to assessing one's own memorizing ability, two different kinds of predictions have frequently been elicited: *Global predictions*, in which people judge how many items of an entire study list they will subsequently recall, and *item-by-item predictions*, which requires people to predict the likelihood of subsequent recall separately for each item (Nelson et al., 1994; Perlmutter, 1978). These latter, so-called judgments of learning (JOLs), are typically elicited by presenting a set of paired associates, for example, Swahili cue words, and English translation equivalents as target words (e.g. Adha – Trouble). After the presentation of each word pair, participants are asked to give a JOL by judging the probability that they will remember the target word a few minutes later when prompted with the cue word.

Metamemory accuracy is typically conceptualized in two ways: One way to define accuracy refers to *relative accuracy or resolution* (Koriat et al., 2002; Nelson & Dunlosky, 1991), which is commonly indexed by an average within-participant gamma correlation

[11] I gratefully acknowledge the help of Daniel Zimprich in preparing the manuscript.

between JOLs and actual memory performance (Nelson, 1984). By contrast to resolution, *absolute accuracy or calibration* pertains to the correspondence between mean JOLs and mean recall performance in a memory test (see Metcalfe, 1998). In both experiments described in the present paper, we will focus on absolute accuracy measures because we aimed at examining the underconfidence-with practice effect which will be described in some detail later and which is reported for this type of measure only (Koriat, 1997).

Findings from studies using relative and absolute accuracy indicate that the ability to predict one's own memory performance is moderate (Koriat, 1997; Koriat et al., 2002; Mazzoni & Nelson, 1995; Scheck et al., 2004; Schneider et al., 2000). These findings can be generalized across different age groups. A characteristic regarding older adults memory predictions, however, seems to be the finding that they "overpredict" their own memory performance (Bruce et al., 1982; Mazzoni & Nelson, 1995; Schneider et al., 2000). Although some researchers have found that older adults' predictions are accurate (e.g., McDonald-Miszczak et al., 1994; Rebok & Balcerak, 1989), the predominant finding is that older adults overestimate their memory performance compared to younger adults (Bruce et al., 1982; Coyne, 1985; Devolder, Brigham, & Pressley, 1990; Murphy, Sanders, Gabriesheski, & Schmitt, 1981; Perlmutter, 1978; Rebok & Balcerak, 1989). Lovelace (1990) hypothesized that, in relation to younger adults, older persons are more prone to prediction errors, generally in the direction of overestimating memory performance because older adults may actually be expecting or demanding more of the memory system than younger adults are.

It is, however, possible to raise the relative and absolute accuracy of JOLs in young and old substantially by eliciting JOLs with a certain *delay* after the presentation of the paired associates. The first to report this effect were Nelson and Dunlosky (1991), who asked participants to memorize paired associates. After a given delay (between 10 and 33 items), respondents were prompted with the cue word and were asked to give a JOL. Both, the absolute and relative accuracy of these delayed JOLs were found to be superior as opposed to that of immediate JOLs. Possibly, when JOLs are delayed, participants rely more heavily on cues pertaining to the ease with which the target can be retrieved from memory and, hence, the accuracy of predicting the recall probability is enhanced, that is, a delayed JOL instantiates a first recall attempt which offers the respondent a first impression of the subjective item difficulty (Dunlosky & Nelson, 1994; Koriat, Ma'ayan, Sheffer, & Bjork, 2006). This delayed-JOLs effect was consistently found in a number of studies and to

practically the same extent in younger and older adults (Connor et al., 1997; Dunlosky & Nelson, 1992; Koriat, Ma'ayan, Sheffer, & Bjork, 2006; Scheck et al., 2004).

Underconfidence-With-Practice effect

Another way to increase the familiarity with the stimulus material is to present the items in more than one learning and recall trial. Giving participants more than one learning occasion leads to an increase in the relative accuracy of judgments, that is, the resolution typically increases over the trial course (Koriat et al., 2002). This finding, however, contrasts with the observation that calibration seems to be impaired with practice. Koriat (1997) reported a discrepancy between JOLs and recall performance which arose with repeated presentation of the stimuli: In two studies participants memorized a list of paired associates in several learn test cycles and, following the study of each pair, provided JOLs. A comparison of the effects of practice on JOLs and actual memory performance, in terms of absolute accuracy, disclosed a pattern the author referred to as the *underconfidence-with-practice (UWP) effect*: While the recall performance increased from the first to the second learning occasion, the effects of practice did not lead to more accurate predictions. Instead of improving calibration, that is, the difference between JOL and actual recall, average JOLs for the second occasion became markedly lower (underconfident) than recall performance. Briefly, UWP thus refers to a loss in accuracy in calibration across practice trials. This is also true for delayed JOLs, in which a smaller, but still relevant UWP effect was reported. In a subsequent review of several studies requiring participants to give JOLs, the UWP effect proved to be robust against a number of experimental manipulations (see, e.g., Koriat et al., 2002). The UWP effect has also been replicated by other authors who investigated JOLs under different conditions (e.g., Meeter & Nelson, 2003; e.g., Serra & Dunlosky, 2005). Investigations of the UWP effect in older adults, however, have not been conducted up to date.

Three theoretical perspectives have been discussed and presented in Chapter 1.3.1, which geared to explain the UWP effect: The cue-utilization framework, the anchoring-and-adjustment effect, and the dual-factors hypothesis. In sum, the three explanatory accounts for the UWP effect appear to be complementary rather than disjunctive, as they describe similar processes from different perspectives. While the anchoring process and the dual-factor hypothesis are mainly geared to explain how JOLs are generated without prior task-related

knowledge or cues about actual recall level, the cue-utilization approach explains the formation of JOLs including feedback from prior learning occasions.

The present study

Viewed from an aging perspective, the reported empirical data does not allow drawing strong conclusions about the UWP effect in older adults or about the impact of a psychological anchor on the formation of JOLs because the available data is mainly restricted to studies examining young adults. Hence, the presence of an UWP effect in absolute accuracy judgments remains to be tested in older adults, whereby, at the same time, we aimed at gauging the UWP effect. There is, however, evidence that the UWP effect observed in young persons is not necessarily replicable in older adults. Connor et al. (1997) pointed out that older persons tend to overestimate their memory performance to a greater extent than younger adults, which may also results from lower recall performance in older persons. Lower recall performance and larger overestimation might influence the UWP effect as well. Note that the UWP effect can only result from JOLs which are markedly lower than the actual recall performance, that is, in order to instantiate the effect, participants are required to underestimate the benefit from additional learning occasions. Older adults, however, recall fewer words and learn less during repeated presentation compared to young adults (Kausler, 1994)

Thus, the overarching goal of the present study was to examine the UWP effect in older persons and provide further empirical data to accuracy judgments in older and younger adults. More specifically, we (I) compared absolute accuracy judgments elicited by younger and older adults for easy and difficult stimulus material, and (II) compared the effects of immediate versus delayed JOLs in both age groups. Finally, we (III) addressed the UWP effect in older adults across two and five study test cycles.

2.3.2 Experiment 1

Experiment 1 was designed to investigate differences in the accuracy of immediate and delayed JOLs between young and older adults using easy and difficult word pairs across two trials.

2.3.2.1 Method

Participants, Design, and Items. Thirty-six young adults (M = 25.7 years, SD = 3.9) and thirty-six older adults (M = 65.8 years, SD = 4.3) from the city of Zurich participated in this study. The experiment was a 2 (JOL timing: Immediate vs. delayed) × 2 (trial: Trial 1 vs. trial 2) × 2 (difficulty: Difficult vs. easy items) × 2 (age group: Young vs. old participants) × 2 (measure: JOL vs. recall performance) design with age group as a between-subjects factor and JOL Timing, Trial, Difficulty, and Measure as within subjects factors. The difficulty of items was determined by combining a German cue word with another German target word (easy: Kitchen - Car) or a Turkish word with its German translation equivalent (difficult: Mesnet - Agency). Note that first the Turkish words were selected to control for word length and syllables and then the German equivalent was matched.

Apparatus and Procedure. The experiment was programmed and executed in Inquisit (Version 1.33) on a Dell personal computer running Microsoft Windows XP Pro system software. Stimuli were presented on a 17" LCD display set at 1024 x 768 pixels.

Participants saw, in the center of the screen, 18 German word pairs and 12 Turkish-German word pairs for 3.5 s each. To rule out position effects, in each presentation cycle items were randomized anew for each participant, within six blocks containing five items (three easy and two difficult items). Half of the easy items and half of the difficult items were randomly allocated to the immediate or to the delayed JOL condition but the process of randomization was manipulated such that words from the first third remained in that third for the second presentation. The same manipulation was applied on words which were presented in the last third. This allocation remained the same across trials. The procedure for the second trial was the same as it was in trial one and word pairs remained in the same JOL-timing condition (immediate vs. delayed). Immediate JOLs were given after the presentation of the cue word by asking participants to answer the following question: "With what probability will you remember the target word in about five minutes from now if you see the cue word? (0 = *will definitely not recall*, 20 = 20% probability of recalling the word, 40 … 100 = *will definitely recall.*"). In the delayed JOLs condition, participants were asked to answer the same query as given for immediate JOLs, but the cue word appeared with a delay of, on average, 45 s after the presentation of the word pair in question. All JOLs were self paced and the participants' response prompted the next item. Finally, a self-paced recall test was administered, where the cue word was presented and the participants were asked to recall the

corresponding target word. After the recall test, the second trial started using the same items. The second presentation cycle was again completed with the self-paced recall test.

2.3.2.2 Results

Mean JOLs (dashed line) and mean recall levels (solid line) are depicted in Figure 2.4, where each JOL timing condition (immediate vs. delayed) paired with each item difficulty (easy vs. difficult) is plotted in four panels. Further, the two age groups are represented by gray (young) and black (old) lines. Over- and underconfidence, defined by the difference between JOLs and recall level, is shown in Figure 2.5 using the same scheme as in Figure 2.4. Overconfidence is represented by positive bars and underconfidence is represented by negative bars. In Table 2.6 means and standard deviations of JOLs and recalled words are reported. In what follows, we address each condition in turn.

Anchoring. In order to test for the presence of an anchor in forming JOLs during the first trial, we compared JOLs from both age groups in all four conditions. Both groups were compared by independent *t*-tests in all four conditions.

In the *Easy × Immediate* condition, both age groups showed comparable JOL levels ($t(70) = 0.47$, $p > .05$, $\eta^2 = .00$, $\Delta JOL_{Young - Old} = 2.04$) which differed from recall. At the same time, the recall performance was significantly lower in the older, compared to the younger group ($t(70) = 2.53$, $p < .05$, $\eta^2 = .08$, $\Delta Recall_{Young - Old} = 9.88$) implying more overconfidence in the older group. A similar pattern was found in the *Difficult × Immediate* condition: The mean JOL level was comparable in both groups ($t(70) = 0.70$, $p > .05$, $\eta^2 = .01$, $\Delta JOL_{Young - Old} = 3.52$) while the recall level was again markedly lower in older participants ($t(70) = 3.76$, $p < .01$, $\eta^2 = .17$, $\Delta Recall_{Young - Old} = 9.88$). The results seemed to support the notion of an anchor determining largely the location of the JOLs in the first trial. To test the assumption that participants were still able to differentiate between easy and difficult items in the immediate condition, an ANOVA with a 2 × 2 design (Difficulty × Age) was calculated. The main effect of difficulty was statistically significant ($F(1, 70) = 46.86$, $p < .01$, $\eta^2 = .40$) while the main effect of age and the interaction of Measure × Age were not statistically significant, indicating that easy items received higher JOLs than difficult items, at approximately the same extent in both age groups at the first trial.

Table 2.6: *Means and standard deviations of JOLs and recalled words in Experiment 1*

Experiment 1

			JOL Timing			
			Immediate		Delayed	
			Easy	Difficult	Easy	Difficult
Young	Trial 1	JOLs	44.0 (18.3)	34.8 (20.2)	29.4 (18.1)	16.8 (15.2)
(n=36)		Recall	22.5 (16.9)	12.0 (14.9)	22.2 (17.4)	6.5 (9.0)
	Trial 2	JOLs	41.6 (21.5)	30.6 (23.9)	44.1 (24.7)	28.9 (21.0)
		Recall	52.5 (24.0)	35.2 (27.4)	52.5 (25.1)	34.9 (22.6)
Old	Trial 1	JOLs	42.0 (18.6)	31.2 (22.6)	17.7 (11.6)	10.4 (9.7)
(n=36)		Recall	12.7 (16.2)	2.2 (5.2)	9.3 (11.1)	2.2 (4.5)
	Trial 2	JOLs	29.8 (23.1)	16.9 (20.0)	19.3 (16.1)	10.6 (12.4)
		Recall	27.2 (26.4)	14.2 (20.8)	22.5 (19.4)	10.8 (14.2)

Note. Standard deviations are in parentheses.

In the *Easy × Delayed* condition the two groups differed in both, the JOL and recall levels (JOL: $t(70) = 3.25$, $p < .01$, $\eta^2 = .13$, $\Delta JOL_{Young - Old} = 11.67$; Recall: $t(70) = 3.77$, $p < .01$, $\eta^2 = .17$, $\Delta Recall_{Young - Old} = 12.96$). Similar results were found in the *Difficult × Delayed* condition (JOL: $t(70) = 2.14$, $p < .05$, $\eta^2 = .06$, $\Delta JOL_{Young - Old} = 6.42$; Recall: $t(70) = 2.59$, $p < .05$, $\eta^2 = .09$, $\Delta Recall_{Young - Old} = 4.32$) where older participants showed both smaller JOLs and lower recall levels than young participants. The ANOVA with the factors difficulty and age yielded significant main effects of difficulty ($F(1, 70) = 49.30$, $p < .01$, $\eta^2 = .41$) and age ($F(1, 70) = 9.16$, $p < .01$, $\eta^2 = .12$), but not a significant interaction of both main effects. Hence, easy items received significantly higher JOLs compared to the difficult items and older participants showed in both difficulty levels lower JOLs than the young participants.

Group analyses. In the following sections, the responses in JOLs and recall of both age groups were analyzed by means of repeated ANOVA and paired *t*-tests.

Easy Immediate: The main effect of trial ($F(1,70) = 18.85$, $p < .01$, $\eta^2 = .21$) was statistically significant, implying higher overall means in trial two, that is, higher means for JOLs and correctly recalled words in both groups compared to trial one. The main effect of measure ($F(1,70) = 18.20$, $p < .01$, $\eta^2 = .21$) was statistically significant as well: On average, means of the JOLs were higher than the average number of correctly recalled words across

both trials and both groups. Further, the young age group generally showed higher means in JOLs and recalled words than the older group, which lead to a statistically significant main effect of age ($F(1,70) = 11.34$, $p < .01$, $\eta^2 = .14$). In terms of effect sizes, both Trial and Measure explained virtually the same amount of the total variance, whereas Age contributed somewhat less to the explanation of the total variance. Relevant for the examination of the UWP effect are the interaction terms: A flatter or opposite slope for the JOLs, compared to the slope of correctly recalled words between the first and the second trial are a prerequisite for the UWP effect to occur. In fact, the two-way interaction effect of Trial × Measure ($F(1, 70) = 117.75$, $p < .01$, $\eta^2 = .63$) was significant. As can bee seen from Figure 2.4, top left panel, the mean JOLs in the first trial are higher than the correctly recalled words. This relationship is inverted in the second trial, where the mean of correctly recalled words is higher than the average JOL. However, it remains to be tested if mean JOLs are significantly lower than mean recall performance. Further, the interaction of Measure × Age ($F(1, 70) = 4.56$, $p < .05$, $\eta^2 = .61$) and the interaction of Trial × Age ($F(1, 70) = 13.52$, $p < .01$, $\eta^2 = .16$) were significant as well, indicating that older participants showed a greater discrepancy between JOLs and recall level and a lower increase in response levels between Trial 1 and Trial 2, compared to the young group. Both interaction terms involving Measure yielded effect sizes over .60 (η^2).

T-tests of the relevant effects yielded different results for young and old participants regarding the UWP effect. This is shown in Figure 2.5, which depicts the differences between mean JOLs and mean recall performance. Compared to the average recall performance, both age groups overestimated the likelihood for recalling easy items receiving immediate JOLs at the first trial (Young: $t(35) = 6.00$, $p < .01$, $\eta^2 = .50$, $\Delta_{JOL\text{-}Recall} = 21.5$; Old: $t(35) = 7.55$, $p < .01$, $\eta^2 = .62$, $\Delta_{JOL\text{-}Recall} = 29.3$). Decisive for the UWP effect is the relation of mean JOLs and mean recall level in trial two. In order to instantiate the effect the mean of confidence judgments must be significantly lower than mean recall. In fact, the UWP effect was found in the younger ($t(35) = -2.67$, $p < .05$, $\eta^2 = .17$, $\Delta_{JOL\text{-}Recall} = -10.9$), but not in the older group ($t(35) = 0.58$, $p > .05$, $\eta^2 = .01$, $\Delta_{JOL\text{-}Recall} = 2.6$) where, on average, JOLs almost matched mean recall performance.

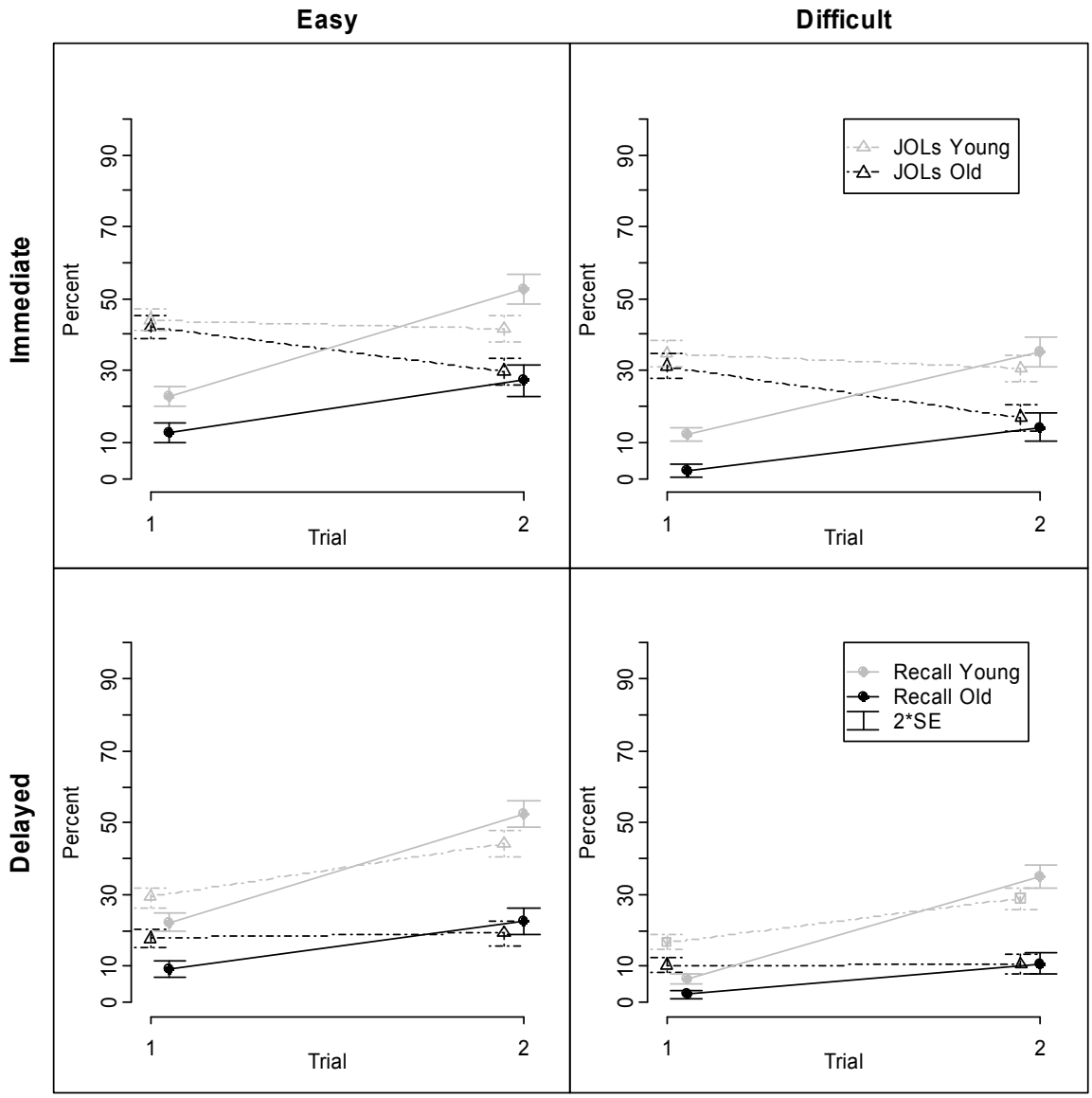

Figure 2.4: Means of JOLs and recalled words are depicted in four panels representing two difficulty conditions and two timing conditions. Hatched lines represent JOLs and solid lines represent recall performance; grey symbolizes young adults and black old adults. The anchoring effect in the immediate condition in trial 1 is apparent upon inspection, where both age groups start at almost identical levels. In the delayed conditions, the JOLs are closer to the recall level and the UWP effect is smaller in young participants.

Difficult Immediate: All three main effects were statistically significant (Trial: $F(1, 70) = 5.98$, $p < .02$, $\eta^2 = .08$; Measure: $F(1, 70) = 25.63$, $p < .01$, $\eta^2 = .27$; Age: $F(1, 70) = 11.86$, $p < .01$, $\eta^2 = .15$) implying that the means of the JOLs and recall performance were higher in the second trial. In addition, the JOLs were, on average, higher than the recall performance, and the young group judged their learning higher and recalled on average more

words than the older group. Both two-way interactions combined with Trial, Trial × Measure ($F(1, 70) = 95.88$, $p < .01$, $\eta^2 = .58$) and Trial × Age ($F(1, 70) = 9.67$, $p < .01$, $\eta^2 = .12$) were statistically significant (see top right panel in Figure 2.4).

In both age groups, the average JOLs were larger than the mean recall performance after the first trial (Young: $t(35) = 6.86$, $p < .01$, $\eta^2 = .57$, $\Delta_{JOL\text{-}Recall} = 22.7$; Old: $t(35) = 7.23$, $p < .01$, $\eta^2 = .60$, $\Delta_{JOL\text{-}Recall} = 29.1$). Effect sizes were comparable to the Easy × Immediate condition, implying that the difficulty of the items did not substantially affect the discrepancy between both measures in the first trial, that is, the JOLs at trial one seemed to be independent of item difficulty. In the second trial, however, the JOLs were not significantly different from the recall performance (Young: $t(35) = -1.08$, $p > .05$, $\eta^2 = .03$, $\Delta_{JOL\text{-}Recall} = -4.6$; Old: $t(35) = 0.63$, $p > .05$, $\eta^2 = .01$, $\Delta_{JOL\text{-}Recall} = 2.7$). Hence, the UWP effect was observed for neither age group in this condition (see top right panel in Figure 2.5).

Easy Delayed: In the condition where easy items received delayed JOLs, the main effects of trial ($F(1,70) = 88.93$, $p < .01$, $\eta^2 = .56$) and of age ($F(1,70) = 28.56$, $p < .01$, $\eta^2 = .29$) were statistically significant but not the main effect of measure ($F(1,70) = 0.52$, $p > .47$, $\eta^2 = .01$). This implied that, on average, the responses given by the participants were higher in the second trial than in the first trial. In addition, the young group had, on average, higher means than the older group which can be seen from Figure 2.4, lower right panel. In addition, the interaction effect of Trial × Measure ($F(1, 70) = 50.88$, $p < .01$, $\eta^2 = .42$) was statistically significant. The interaction can be seen from Figure 2.4, where the increase in mean JOLs from trial one to trial two was not as steep as the increase in correctly recalled words. The interaction term of Age × Trial ($F(1, 70) = 22.51$, $p < .01$, $\eta^2 = .24$) was statistically significant as well, the explained variance, however, was lower compared to the previously described interaction. The increase in the mean responses (JOLs and recalled words) was less pronounced for the older group compared to the younger group indicating a greater change for the younger between both trials.

Next, mean JOLs and recall performance was compared by means of *t*-tests. In Trial one both age groups significantly overestimated the mean of recalled words (Young: $t(35) = 2.51$, $p < .05$, $\eta^2 = .15$, $\Delta_{JOL\text{-}Recall} = 7.2$; Old: $t(35) = 4.31$, $p < .01$, $\eta^2 = .35$, $\Delta_{JOL\text{-}Recall} = 8.5$). Note, however, that effect sizes were considerably smaller than they were in the Easy × Immediate condition.

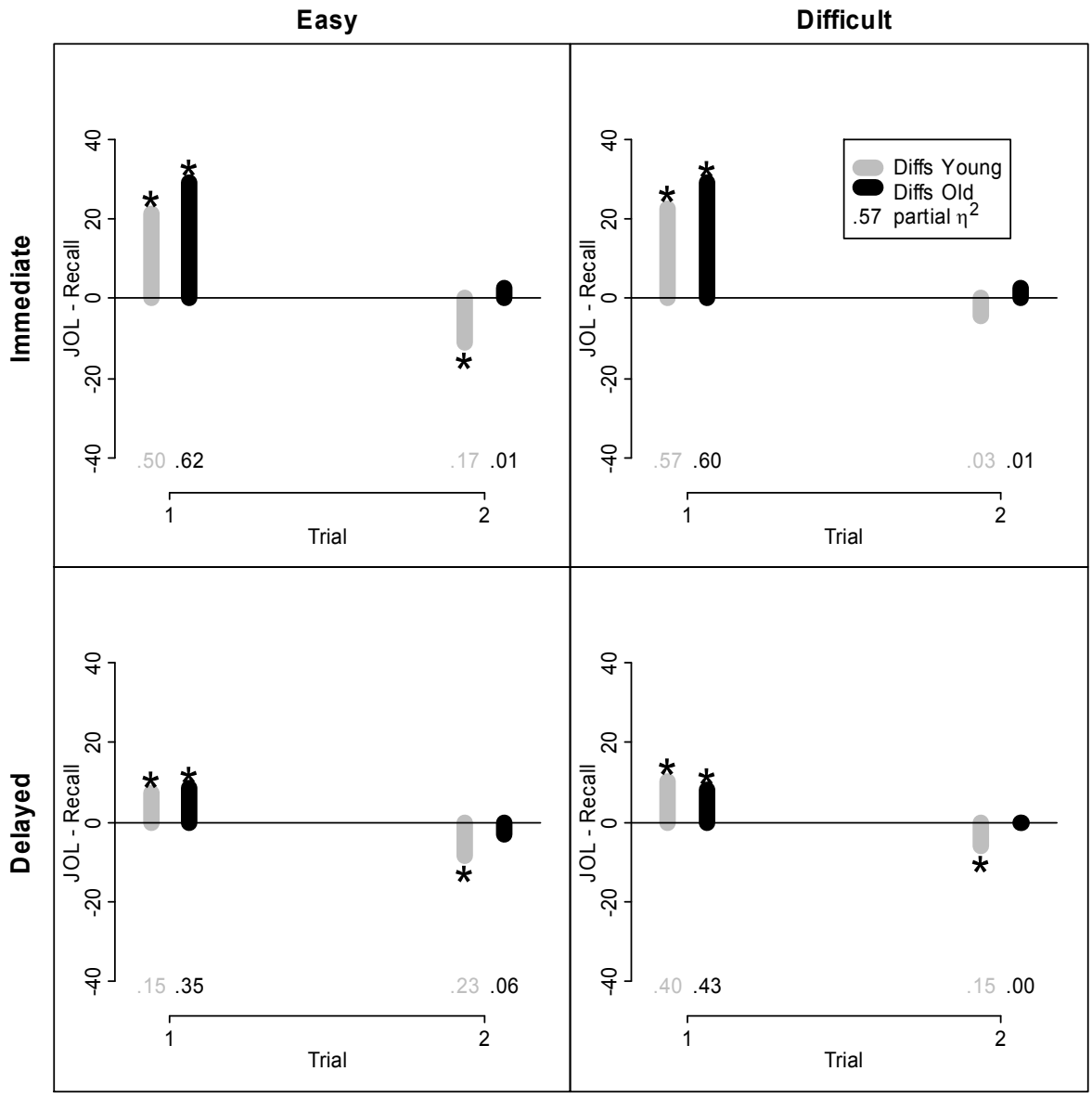

Figure 2.5: In order to make the UWP effect more apparent, the difference between JOLs and recall performance are displayed in all four conditions, represented by grey (young group) and black (old group) bars. Black stars represent statistically significant differences, at the $p < .05$ level, between JOLs and recall performance. The UWP effect is present if in trial 2 the average JOL level is significantly lower than the recall level, which is the case for young participants in both easy, and in the easy x difficult condition.

In Trial two, only the young group underestimated the actually recalled level of the target words ($t(35) = -3.25$, $p < .01$, $\eta^2 = .23$, $\Delta_{JOL\text{-}Recall} = -8.3$), the JOLs of the older persons were, on average, not significantly different from their mean recall performance ($t(35) = -1.47$, $p > .05$, $\eta^2 = .06$, $\Delta_{JOL\text{-}Recall} = -3.2$). Hence, the UWP effect was found for the young group only.

Difficult Delayed: The ANOVA for items receiving delayed JOLs yielded statistically significant main effects (Trial: $F(1, 70) = 65.39$, $p < .01$, $\eta^2 = .48$; Measure: $F(1, 70) = 5.64$, $p < .02$, $\eta^2 = .08$; Age: $F(1, 70) = 24.63$, $p < .01$, $\eta^2 = .26$). The two way interactions of Trial × Measure ($F(1, 70) = 44.47$, $p < .01$, $\eta^2 = .39$) and of Trial × Age ($F(1, 70) = 26.96$, $p < .01$, $\eta^2 = .28$) as well as the triple interaction of Trial × Measure × Age ($F(1, 70) = 4.46$, $p < .05$, $\eta^2 = .06$) were all statistically significant. The significant two-way interaction of Trial × Measure indicates a possible UWP effect due to less steep slopes between the first and the second trial for JOLs compared to the slopes of the correct recall performance (see Figure 2.4, lower left panel).

In Trial one, both age groups overestimated the average recall performance (Young: $t(35) = 4.81$, $p < .01$, $\eta^2 = .40$, $\Delta_{JOL\text{-}Recall} = 10.3$; Old: $t(35) = 5.09$, $p < .01$, $\eta^2 = .43$, $\Delta_{JOL\text{-}Recall} = 8.2$), in Trial two, however, only the younger participants showed the UWP effect ($t(35) = -2.44$, $p < .01$, $\eta^2 = .15$, $\Delta_{JOL\text{-}Recall} = -6.0$). For the older participants, on average, the JOLs were virtually identical to the mean of actually recalled words ($t(35) = -0.09$, $p > .05$, $\eta^2 = .00$, $\Delta_{JOL\text{-}Recall} = -0.2$).

2.3.2.3 Discussion

Experiment 1 was designed to investigate and compare the UWP effect in younger and older adults. The young and the older group showed comparable *immediate* JOLs in the *first trial* within both difficulty levels, hence, analogous to the findings from Scheck and Nelson (2005), mean JOLs were within the range of a psychological anchor of 30% to 50% correct recall. At the same time, the recall level in both groups was lower than 30%, which lead to substantially overconfident JOLs. Note that older adults recalled markedly fewer items than young participants and, consequently, the predictions of the older participants were more overconfident than those of the young. These findings are in line with results from earlier studies where older adults accuracy judgments were more biased toward overconfidence than those of the younger adults (Connor et al., 1997; Murphy, Sanders, Gabriesheski, & Schmitt, 1981; Touron & Hertzog, 2004). In sum, the JOLs in the first trial across both groups seemed to be influenced by a psychological anchor because these judgments appeared to be unaffected by actual recall performance, that is, it appears that if the respondents have no information about item difficulty they seem to rely on an anchor when giving JOLs (Scheck & Nelson, 2005).

For *delayed* JOLs in the *first trial*, we expected JOLs to be more accurate (Nelson & Dunlosky, 1991). In fact, participants were able to give more precise JOLs compared to the immediate condition. Accordingly, JOLs differed between both age groups and also across both difficulty conditions, that is, other than immediate JOLs, delayed JOLs were influenced by item difficulty. Even though JOLs were closer to the recall level, they still did not predict the recall probability correctly. Hence, delayed JOLs seemed to rely on monitoring processes, which delivered more precise informations about the probability of recalling a specific item, *and* on anchoring mechanisms, which upward-biased the judgements to some degree. These findings fit in well with the results from Scheck et al. (2004) who investigated the dual-factors hypothesis and reasoned that immediate JOLs are based on an anchor while delayed JOLs rely also on informations about item difficulty stemming from monitoring processes.

In the *second trial*, the accuracy of JOLs seemed to be boosted by the additional learning trial with JOLs being fairly close to the actually recalled words. The increasing influence of monitoring on the formation of *immediate* JOLs relative to that of an anchor most probably enhanced the accuracy of judgments. However, there was a difference regarding the pattern of accuracy judgments between both age groups: The young group displayed the UWP effect for easy word-pairs, as described by Koriat (1997), even though the effect size of the underconfident judgements was just a third compared to the overconfidence effect reported in the first trial. The older group, in turn, predicted the recall level correctly. Hence, items given in the second trial were judged by both groups more precisely but older participants seemed to be more successful in rating their recall performance accurately. In the *delayed* condition, a similar pattern was observed for older participants. The second learning trial increased the accuracy of JOLs, which then matched the actual recall level. The young group, in comparison, did not benefit to the same extent from the additional presentation. For both difficulty levels, the young displayed the UWP effect, that is, the overconfidence from Trial one turned into underconfidence in Trial two. This effect was greater for easy items than it was for difficult items. An apparent difference between both groups was the level of memory performance, also in terms of learning effects across both trials: Apart from the higher initial level, younger participants seemed to benefit more from the additional learning trial than older participants. As a consequence, older participants' performance did not exceed the level expected from their JOLs.

Thus, with respect to the UWP effect, the results from both age groups were disparate. Young participants showed an UWP effect in most conditions, except for immediate JOLs given for difficult word pairs. Overall, the UWP effect in the young group can be explained by anchoring *and* monitoring processes (Koriat et al., 2002; Scheck & Nelson, 2005). Results from the older group yielded a different pattern: Both age groups overestimated their performance in the first trial, but the older group differed from the young group mainly in the second trial, where JOLs given by older participants matched, on average, the correctly recalled words. In fact, the UWP effect was not observed in any of the four conditions in older participants. The complete absence of the UWP effect in the older age group appears remarkable when considering that the effect has been shown to withstand several manipulations and, thus, was thought to be very robust (see Koriat et al., 2002). Several reasons may have contributed to the non-occurrence of the effect in the older group: In the *immediate* condition, the initial overconfidence in older participants was larger than in the young group which, in turn, implies that older participants would have needed to downgrade their judgements in the second trial to a larger amount than young participants in order to underestimate their performance in the second trial. In fact, the older group showed an interaction effect between JOLs and recall performance and downgraded substantially their JOLs in the second trial, but probably the relatively flat learning trajectory prevented judgments from being underconfident. In the *delayed* condition, again, older adults were almost perfect at predicting their recall performance. As in the immediate condition they showed a very flat learning trajectory which may have contributed to the absence of the UWP effect. What distinguished older adults clearly from younger adults was the smaller benefit in learning performance resulting from the second trial, which lead to a low mean recall performance that is problematic in light of the UWP effect: If recall performance is low, *under*confidence in JOLs can hardly be achieved. A straightforward approach to deal with the problem of low mean performance in recalled words in older participants would be to increase the number of learning occasions in order to raise the recall performance above the JOL level and finally elicit the UWP effect also in older adults. As a consequence, we conducted a second experiment by administering five, instead of two, learning trials.

2.3.3 Experiment 2

Experiment 2 can be seen as an expansion of Experiment 1. Hence, it was mainly designed to maintain the first two trials comparable with Experiment 1 but give participants the possibility to learn and recall the paired associates in five trials – instead of two. The primary aim was test the assumption made in Experiment 1 that the absence of the UWP effect in the older group stems from too few presentation cycles, that is, older participants may need more than two learning trials until their recall level exceeds their JOLs. We expected the UWP effect in younger participants to be unaffected by extending the number of learning trials.

The general hypotheses remain the same as in Experiment 1: If the recall level is lower than the anchor, overconfidence is expected in the first trial. For each consecutive trial, we expect an increasing approximation of JOLs and previous recall performance. In older adults, underconfidence, and the UWP effect, may result after the second trial.

2.3.3.1 Method

Participants, Design, and Items. Thirty-four young (M = 26.1 years, SD = 3.1) and thirty-four older persons (M = 68.6 years, SD = 3.6) participated in Experiment 2. The design of the study was basically the same as in Experiment 1 with the difference that participants were given five learning occasions instead of two. This lead to a 2 (JOL timing: Immediate vs. delayed) × 5 (trial: Trial 1 to trial 5) × 2 (difficulty: Difficult vs. easy items) × 2 (age group: Young vs. old participants) × 2 (measure: JOL vs. recall performance) design.

Apparatus and Procedure. The apparatus and the procedure used here were the same as in Experiment 1. To increase precision of measurement, the number of stimuli was doubled to 36 German word pairs and 24 German – Turkish word pairs. Hence, the randomization procedure was extended to five presentation cycles consisting of six blocks containing ten items (six easy and four difficult items).

2.3.3.2 Results

As in Study 1, a repeated measures ANOVA with the between-subject factor age (young vs. old group) was computed. Here, participants were given five trials which lead to a 5 × 2 × 2 (trial, measure, age group) design.

Table 2.7: *Means and standard deviations of JOLs and recalled words across both experiments*

Experiment 2

			JOL Timing			
			Immediate		Delayed	
			Easy	Difficult	Easy	Difficult
Young	Trial 1	JOLs	47.8 (21.0)	33.4 (22.0)	36.4 (20.5)	17.2 (18.4)
(n=34)		Recall	28.6 (16.6)	8.1 (10.9)	26.1 (18.5)	8.3 (10.9)
	Trial 2	JOLs	53.2 (20.3)	32.1 (17.4)	51.5 (21.5)	31.4 (18.5)
		Recall	60.1 (21.2)	38.9 (22.3)	55.4 (23.2)	32.1 (21.1)
		Trial 3	JOLs	72.3 (18.7)	54.5 (21.3)	67.8 (20.6)
	Recall	75.3 (20.2)	66.9 (19.6)	72.2 (20.2)	60.0 (23.9)	55.8 (24.4)
		Trial 4	JOLs	81.3 (17.1)	75.5 (18.8)	78.2 (18.3)
		Recall	84.5 (16.0)	82.8 (19.1)	82.0 (17.1)	71.7 (22.9)
	Trial 5	JOLs	88.3 (12.6)	85.9 (16.1)	85.6 (14.0)	77.9 (19.2)
		Recall	89.4 (11.8)	90.4 (14.1)	88.1 (13.6)	81.4 (19.1)
Old	Trial 1	JOLs	48.6 (26.9)	38.4 (31.3)	23.6 (22.8)	85.3 (17.5)
(n=34)						
		Recall	7.8 (8.7)	0.7 (2.4)	6.0 (7.7)	14.9 (21.2)
	Trial 2	JOLs	36.5 (27.6)	26.3 (30.7)	23.7 (20.0)	0.5 (2.0)
		Recall	22.9 (16.7)	7.4 (11.0)	20.4 (14.1)	12.5 (13.8)
	Trial 3	JOLs	45.7 (29.2)	30.5 (30.6)	32.7 (22.9)	6.4 (8.5)
		Recall	34.3 (19.3)	15.2 (18.5)	31.4 (19.5)	17.5 (18.0)
	Trial 4	JOLs	52.2 (29.5)	37.8 (33.4)	37.9 (25.4)	12.0 (13.5)
		Recall	42.6 (22.0)	24.3 (24.2)	39.2 (21.6)	22.6 (21.2)
	Trial 5	JOLs	55.9 (29.2)	43.8 (33.0)	42.3 (25.7)	21.3 (18.4)
		Recall	47.2 (21.5)	34.1 (26.6)	44.1 (20.6)	29.6 (26.3)

Note. Standard deviations are in parentheses.

The mean JOLs and the mean recall performance are shown in Figure 2.6 and the difference between JOLs and the recalled words is illustrated in Figure 2.7.

First, we start with combined analyses to examine whether an anchor is affecting JOLs in the first trial. In Table 2.7 mean JOLs and recalled items are reported with standard deviations in parentheses. Then we turn to the group analyses in each condition to examine the UWP effect and the trajectories of JOLs and recall levels in both age groups.

Anchoring. Independent *t*-tests in the *Easy* × *Immediate* condition yielded almost identical JOL levels in both age groups ($t(66) = -0.14$, $p > .05$, $\eta^2 = .00$, $\Delta JOL_{Young - Old} = -0.82$) but a significantly lower recall level in older adults ($t(66) = 6.46$, $p < .01$, $\eta^2 = .39$, $\Delta Recall_{Young - Old} = 20.78$). For *difficult* items judged immediately a similar pattern was found. The JOLs between both groups did not differ significantly ($t(66) = -0.75$, $p > .05$, $\eta^2 = .01$, $\Delta JOL_{Young - Old} = -4.95$), on the same time, older adults recalled substantially fewer words than the young group ($t(66) = 3.82$, $p < .01$, $\eta^2 = .18$, $\Delta Recall_{Young - Old} = 7.35$). These results corroborated the findings from Experiment 1 and supported the notion of a psychological anchor which biases responses in the first trial. Further, it was tested if participants were able to differentiate between difficult and easy items. As in Experiment 1, an ANOVA with a 2 × 2 design (Difficulty × Age) yielded a significant main effect of difficulty ($F(1, 66) = 63.33$, $p < .01$, $\eta^2 = .49$). The main effect of age and the interaction of Difficulty × Age were not statistically significant, indicating that participants in both age groups judged the probability of recalling difficult items lower than the probability of recalling easy items.

In the *Easy* × *Delayed* condition the age groups differed both, in mean JOLs ($t(66) = 2.44$, $p < .05$, $\eta^2 = .08$, $\Delta JOL_{Young - Old} = 12.84$) and in mean recall ($t(66) = 5.85$, $p < .01$, $\eta^2 = .34$, $\Delta Recall_{Young - Old} = 20.10$), indicating that older adults JOLs and recall performance was significantly lower than those from the young group. These findings replicated largely the results from Experiment 1 which signify that, besides anchoring, monitoring processes are also involved in the formation of delayed JOLs. In the *Difficult* × *Delayed* condition, however, the JOLs did not differ between the age groups ($t(66) = 0.48$, $p > .05$, $\eta^2 = .00$, $\Delta JOL_{Young - Old} = 2.30$) although older adults recalled significantly fewer words compared to the young group ($t(66) = 4.14$, $p < .01$, $\eta^2 = .21$, $\Delta Recall_{Young - Old} = 7.84$). The ANOVA yielded a significant main effect of difficulty ($F(1,66) = 93.15$, $p < .01$, $\eta^2 = .59$) and a significant interaction of Difficulty × Age ($F(1,66) = 13.23$, $p < .01$, $\eta^2 = .17$) indicating that participants differentiated between difficult and easy items, by giving easy items higher JOLs.

The young group, however, appeared to react more sensitively on the increased difficulty of items by lowering the mean JOLs to a greater extent compared to the old group.

Group analyses. In each condition, JOLs and recall levels were analyzed by means of repeated measures ANOVA.

Easy Immediate: The main effects for easy word pairs receiving immediate JOLs were all significant (Trial: $F(4, 264) = 208.99$, $p < .01$, $\eta^2 = .76$; Measure: $F(1,66) = 14.08$, $p < .01$, $\eta^2 = .18$; Age: $F(1, 66) = 55.28$, $p < .01$, $\eta^2 = .46$). As can be seen from Figure 2.6, the mean JOLs and the average recall performance increased from trial to trial. The level of the mean responses (JOLs and recall performance) of the young group was, on average, higher as the one from the old group and, in addition, the main effect of measure was significant, because the participants overestimated their performance markedly in the first trial. Older participants judged, on average, their recall performance lower and remembered on average fewer words than the younger group did. As can be seen from Figure 2.6, the greatest changes in level and slope of the mean JOLs and the mean recall performance occurred between trial one and trial two. The interaction effect of Measure (JOL vs. recall performance) and Trial was statistically significant ($F(4, 264) = 53.31$, $p < .01$, $\eta^2 = .45$) indicating different trajectories for mean JOLs compared to mean recall, which is a prerequisite for the UWP effect. Furthermore, the two-way interaction of Trial × Age ($F(4, 264) = 27.66$, $p < .01$, $\eta^2 = .30$) as well as the interaction of Measure × Age ($F(1, 66) = 11.10$, $p < .01$, $\eta^2 = .14$) were statistically significant.

The next step was to compare the mean JOLs with the mean recall performance in each trial, differentiated by age group, to verify if the significant interaction term of Measure × Trial was due to the UWP effect. Figure 2.7, top left panel, offers an overview of the mean discrepancy between JOLs minus correctly recalled words across all four Timing × Difficulty conditions for both age groups. Effect sizes were largest after the first two trials, and then an attenuation of the effects occurred across the five trials. This is also a typical finding from research on verbal learning, that is, the return in recall performance from every additional trial is continually diminishing, leading to growth curves with an asymptotic trajectory (Zimprich, Rast, & Martin, in press).

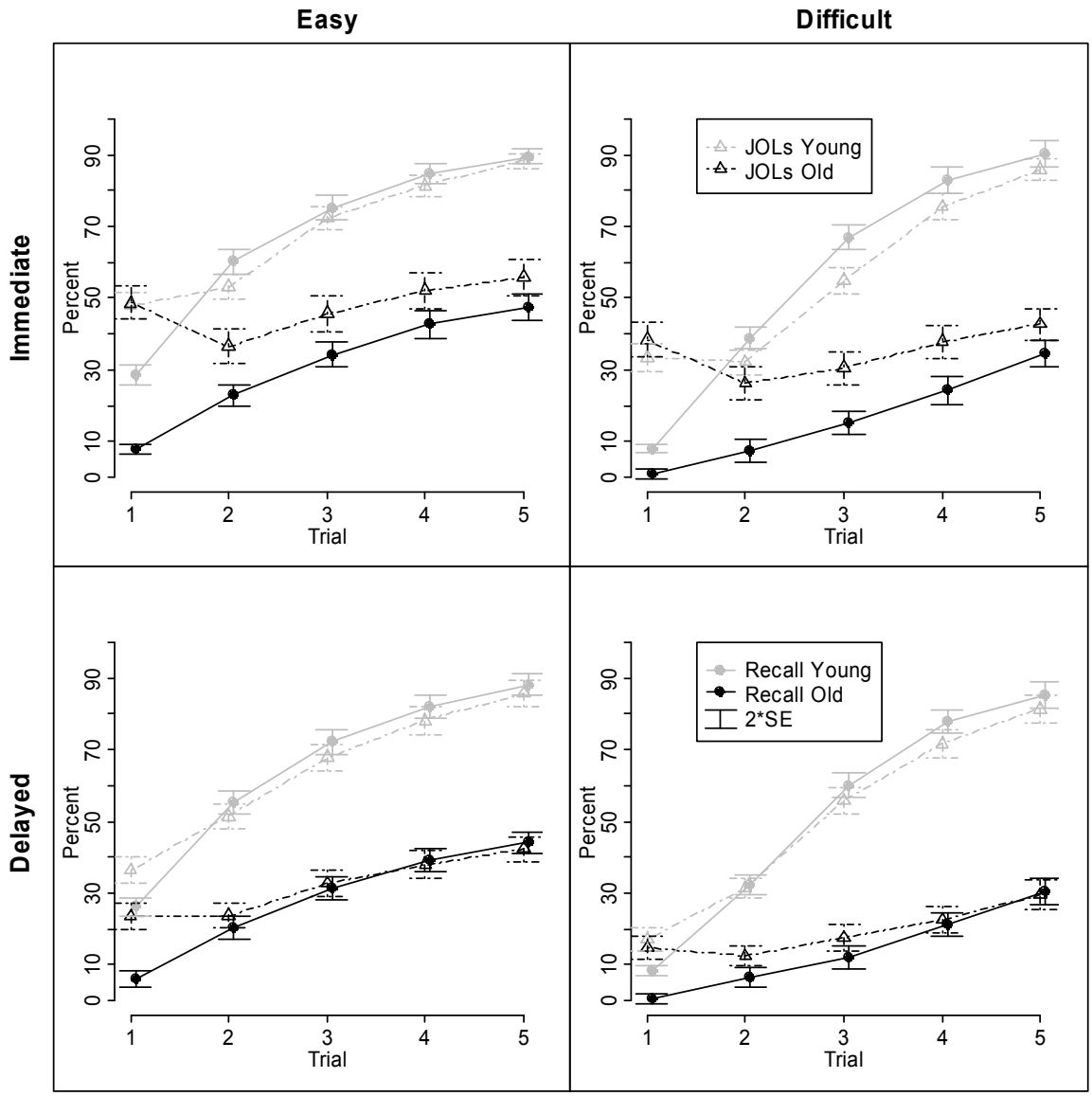

Figure 2.6: Means of JOLs and recalled words from Experiment 2 with five trials across all four conditions. Hatched lines represent JOLs and solid lines represent recall performance. The learning curves for older participants (black) are markedly lower compared to the young group (grey). Nonetheless, the average JOL in the first trial is practically identical in both groups for immediate JOLs which corroborates the notion of an anchoring mechanism. Delayed JOLs are much more accurate and the anchoring effect is smaller in the first trial compared to immediately elicited JOLs.

The young group showed a statistically significant overconfidence in the first trial ($t(33) = 4.73$, $p < .01$, $\eta^2 = .40$, $\Delta_{JOL\text{-}Recall} = 19.2$) and underconfident judgments in the second trial ($t(33) = -2.27$, $p < .05$, $\eta^2 = .14$, $\Delta_{JOL\text{-}Recall} = 6.9$) which indicated the presence of the UWP effect. From trial three on, the mean JOLs were not statistically different from the mean of recalled words anymore.

The older group showed a large and statistically significant overconfidence effect in the first trial ($t(33)= 8.52$, $p < .01$, $\eta^2 = .69$, $\Delta_{JOL\text{-}Recall} = 40.8$) which persisted in trial two ($t(33)= 2.51$, $p < .05$, $\eta^2 = .16$, $\Delta_{JOL\text{-}Recall} = 13.7$), three ($t(33)= 2.28$, $p < .05$, $\eta^2 = .14$, $\Delta_{JOL\text{-}Recall} = 11.4$), and five ($t(33)= 2.10$, $p < .05$, $\eta^2 = .12$, $\Delta_{JOL\text{-}Recall} = 8.7$). Older adults thus seemed to overestimate their recall performance across most of the trials. The effect sizes emphasized this finding with a large effect for the first trial and smaller effect sizes for trials two to five. The UWP effect was not found in the older group at all.

Easy Delayed: The ANOVA for easy word pairs receiving delayed judgments yielded significant main effects of trial ($F(4, 264) = 242.74$, $p < .01$, $\eta^2 = .79$) and of age ($F(1, 66) = 82.31$, $p < .01$, $\eta^2 = .56$). Due to lower JOLs in the first trial, the main effect of measure was not statistically significant (see Figure 2.6, lower left panel). The interactions with Trial (Trial × Measure: $F(4, 264) = 18.53$, $p < .01$, $\eta^2 = .22$; Trial × Age: $F(4, 264) = 25.23$, $p < .01$, $\eta^2 = .28$) were both statistically significant indicating a possible UWP effect and different trajectories across the five trials for both age groups.

In the following, paired *t*-tests for each trial, computed separately for both age groups, are reported. As can be seen from Figure 2.7 (lower left panel), the young group significantly overestimated the recall performance after the first trial ($t(33) = 2.85$, $p < .01$, $\eta^2 = .20$, $\Delta_{JOL\text{-}Recall} = 10.3$), but underestimated the performance in trials three ($t(33) = -2.42$, $p < .05$, $\eta^2 = .15$, $\Delta_{JOL\text{-}Recall} = -4.4$) and four ($t(33) = -2.47$, $p < .05$, $\eta^2 = .16$, $\Delta_{JOL\text{-}Recall} = -3.9$). Compared to the preceding (Easy × Immediate) condition the effect size in trial one was half as large which implies that the discrepancy between JOLs and recall performance was attenuated by the delayed judgments. In the young group, the criterion for an UWP effect was met in the third trial. The older group, on average, also overestimated the recall performance after the first trial ($t(33) = 4.18$, $p < .01$, $\eta^2 = .35$, $\Delta_{JOL\text{-}Recall} = 17.5$). The effect size was half as large as in the preceding immediate condition which indicated that delayed judgments were more precise than immediate. In the consecutive trials, however, mean JOLs all matched mean recall performance which was not compatible with the requirements for the UWP effect. Again, the young group showed the UWP effect, but not the old group.

Difficult Immediate: The main effects for difficult word pairs receiving immediate JOLs were all statistically significant (Trial: $F(4, 264) = 193.60$, $p < .01$, $\eta^2 = .75$; Measure: $F(1, 66) = 10.81$, $p < .01$, $\eta^2 = .14$; Age: $F(1, 66) = 67.95$, $p < .01$, $\eta^2 = .51$). The same was

true for the interaction terms: All two-way interactions (Trial × Measure: $F(4, 264) = 43.91$, $p < .01$, $\eta^2 = .40$; Trial × Age: $F(4, 264) = 60.75$, $p < .01$, $\eta^2 = .48$; Measure × Age: $F(1, 66) = 13.77$, $p < .01$, $\eta^2 = .17$) as well as the tree-way interaction were statistically significant (Trial × Measure × Age: $F(4, 264) = 3.14$, $p < .05$, $\eta^2 = .05$).

Young participants overestimated the performance after the first trial ($t(33) = 5.76$, $p < .01$, $\eta^2 = .50$, $\Delta_{JOL\text{-}Recall} = 27.5$) and, like in the previously described Easy × Immediate condition, they underestimated their recall performance in the second trial ($t(33) = -2.10$, $p < .05$, $\eta^2 = .12$, $\Delta_{JOL\text{-}Recall} = 6.8$). Underconfidence persisted throughout trial three ($t(33) = -4.14$, $p < .01$, $\eta^2 = .34$, $\Delta_{JOL\text{-}Recall} = 12.4$), four ($t(33) = -3.05$, $p < .01$, $\eta^2 = .22$, $\Delta_{JOL\text{-}Recall} = 7.4$) and five ($t(33) = -2.09$, $p < .05$, $\eta^2 = .12$, $\Delta_{JOL\text{-}Recall} = 4.5$). Accordingly, the criteria for the UWP effect was met and the underconfidence grew largest in trial three. The effect size in the first trial was larger compared to the preceding delayed condition indicating a dependency on the timing of the JOL, that is, when JOLs were given immediately, the effect size of the difference between JOLs and recall performance in the first trial was markedly higher compared to the delayed condition. The older group overestimated its performance in the first four trials (Trial 1: $t(33) = 7.07$, $p < .01$, $\eta^2 = .60$, $\Delta_{JOL\text{-}Recall} = 37.6$; Trial 2: $t(33) = 3.13$, $p < .01$, $\eta^2 = .23$, $\Delta_{JOL\text{-}Recall} = 18.9$; Trial 3: $t(33) = 2.59$, $p < .05$, $\eta^2 = .17$, $\Delta_{JOL\text{-}Recall} = 15.3$; Trial 4: $t(33) = 2.43$, $p < .05$, $\eta^2 = .15$, $\Delta_{JOL\text{-}Recall} = 13.6$), in the fifth trial, however, the JOLs did not differ from the actual recall performance any more. The amount of the difference between mean JOLs and the mean of correctly recalled words diminished from trial to trial, up to trial five where it was not significant any more. The same was true for the effect sizes, being the largest in the first trial, and continually diminishing in the consecutive three trials. The older group increased its absolute accuracy with every additional trial and did not display the UWP effect.

Difficult Delayed: In the fourth condition, where difficult word pairs received delayed JOLs, the main effect of trial ($F(4,264) = 202.73$, $p < .01$, $\eta^2 = .75$) and of age ($F(1,66) = 108.50$, $p < .01$, $\eta^2 = .62$) were both statistically significant. The two-way interactions were all statistically significant (Trial × Measure: $F(4, 264) = 12.46$, $p < .01$, $\eta^2 = .16$; Trial × Age: $F(4, 264) = 61.23$, $p < .01$, $\eta^2 = .48$; Measure × Age: $F(1, 66) = 4.93$, $p < .05$, $\eta^2 = .07$), but not the three-way interaction. In Figure 2.6 the reported interactions are depicted in the lower right panel.

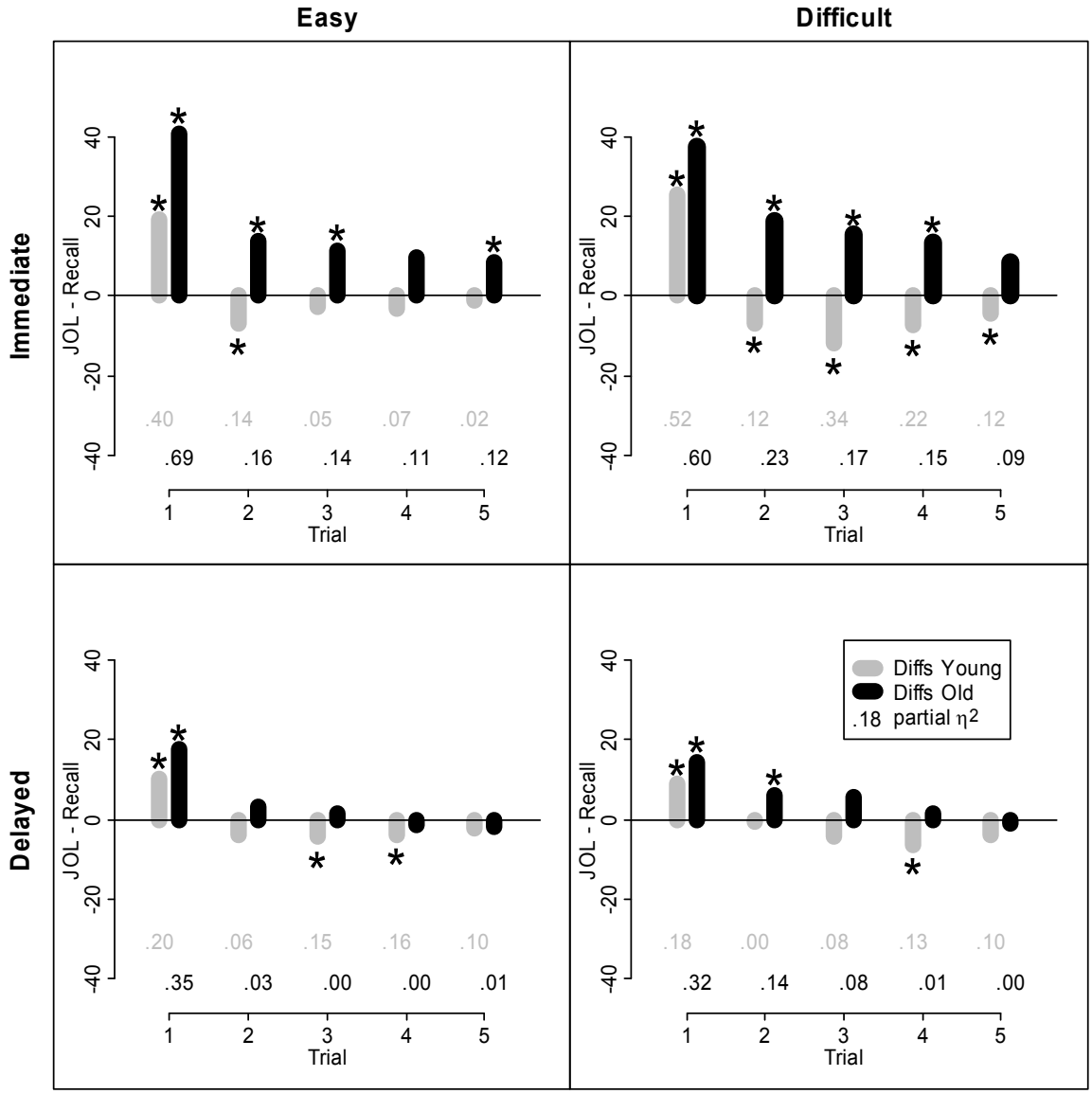

Figure 2.7: The difference between mean JOLs and mean recall performance is represented by black and grey bars. Black stars represent statistically significant differences, at the $p < .05$ level, between JOLs and recall performance. Older adults did not show an UWP effect in any of the four conditions. Young adults, in contrast, displayed the UWP effect in the second trial, for immediately elicited JOLs and in the third or fourth trial for delayed judgments.

As in all previous conditions, both age groups overestimated significantly their recall performance after the first trial (Young: $t(33) = 2.68$, $p < .05$, $\eta^2 = .18$, $\Delta_{JOL\text{-}Recall} = 8.8$; Old: $t(33) = 3.92$, $p < .01$, $\eta^2 = .32$, $\Delta_{JOL\text{-}Recall} = 14.4$). In addition, the young group underestimated their performance in the fourth trial ($t(33) = -2.18$, $p < .05$, $\eta^2 = .13$, $\Delta_{JOL\text{-}Recall} = -6.2$), in all other trials, however, the mean JOLs were not significantly different from the mean of

correctly recalled words: An UWP effect in the young group was not detected in this condition. Apart from the first trial, the older group overestimated the recall performance in the second trial as well ($t(33) = 2.28$, $p < .05$, $\eta^2 = .14$, $\Delta_{JOL\text{-}Recall} = 6.1$). The older group, once again, did not show the UWP effect. As noted before, the effect sizes were smaller, compared to the condition where JOLs were given immediately indicating that delaying judgment of learning enhances absolute accuracy.

2.3.3.3 Discussion

Experiment 2 was designed to investigate and compare the trajectories of JOLs and recall performance in two age groups across five learning occasions. In Experiment 1 we hypothesized that older adults would need more than two learning trials to achieve a recall level that might instantiate the UWP effect, hence, we required participants to complet five learning occasions. The difficulty for both types of items (easy and difficult) was comparable to the difficulty in Experiment 1. Note, however, that the amount of items in Experiment 2 was doubled to a total of 36 easy and 24 difficult word pairs to avoid ceiling effects.

As in Experiment 1, in the *immediate* condition both age groups started out at almost identical JOL levels, while their average recall performance was very different which lead to substantial overconfidence in JOLs. In the first trial the mean JOLs were within 30% to 50% predicted recall, which replicated findings from Experiment 1 and from earlier studies (Scheck & Nelson, 2005) and substantiated the anchoring hypothesis that the magnitude of mean JOLs in the first trial depends largely on a psychological anchor if no prior information about the item difficulty is available. At the same time, the larger overconfidence in JOLs of older adults probably stemmed from the anchoring mechanism, that is, if young and old participants judged items almost independently of their difficulty, this led to larger overconfidence in older adults simply because their recall level is lower.

In the first trial of the *delayed* condition we also expected to find results similar to Experiment 1. In fact, delayed JOLs were generally more accurate compared to the immediate condition. Accordingly, JOLs given for easy items by older adults were smaller than JOLs from young participants. This again substantiated the notion of enhanced accuracy in delayed judgments, probably due to monitoring processes, which was also reported in Experiment 1 and in earlier studies (Koriat, 1997; Nelson & Dunlosky, 1991). An exception seemed to be JOLs given for difficult items, where both age groups had comparable levels.

Regarding the UWP effect, in the young group the results from Experiment 1 were only in part replicated. Participants showed the UWP effect in the Easy × Immediate condition but, contrary to the previous experiment, they also underestimated the difficult words in the Difficult × Immediate condition, which corroborates findings from Scheck and Nelson (2005). In the delayed condition, results from the first two trials did not elicit the UWP effect neither for easy nor for difficult word pairs. Hence, when considering only the first two trials, the UWP effect was found in the immediate but not in the delayed condition. However, if all five trials are taken into consideration, underconfidence can also be observed for word pairs in the delayed condition (see Figure 2.7, Easy × Delayed and Difficult × Delayed). Note that young participants were overconfident only in the first trial. From the second trial on, the mean JOLs remained underconfident or correct throughout consecutive trials.

In the old group, JOLs elicited in the last three trials still did not underscore the recall level and, hence, the UWP effect could not be found: As in Experiment 1, the older group did not show the UWP effect in any of the four conditions and in none of the five learning trials. Our assumptions that older adults would need more trials to display the UWP effect could not be verified in Experiment 2, in this respect, adding three trials did not lead to the expected underconfident judgments in later trials. For immediate JOLs, older participants overestimated their recall performance in most of the trials. Apart from the initial and large overconfidence, in the consecutive four trials older participants seemed to overrate their performance by a rather stable amount. When giving delayed judgments, the estimated recall probability was generally closer to the actual recall performance compared to immediate JOLs. Contrary to the young group, the older participants never underestimated their recall level. Even at the second trial, where the UWP effect typically is found, older adults still overestimated or judged correctly their recall level. Note, however, that the recall performance in the delayed × difficult condition was very low in the first trials. Hence, overconfidence might also be in part due to a floor effect in recall performance, that is, recall levels lower than 10% almost inevitably led to overconfident judgments, because participants hardly downgraded their judgments lower than 10%.

To sum up, in both groups the greatest changes in JOLs took place in the first two trials. Independent of item difficulty or timing of the JOL, the overconfidence effect was always largest in the first trial, followed by a marked decrease in the second trial. This

corroborates the notion of a strong influence of an anchor in the first trial which decreases in favor of monitoring in consecutive trials. In the second trial, young participants downgraded JOLs to underconfidence but older adults adjusted JOLs to be slightly overconfident or correct. In the consecutive three trials the trajectories of both measures (JOLs vs. correct recall performance) appeared to converge, that is, the absolute accuracy increased across all trials. Note that the shape of the subjective accuracy trajectories was comparable across both age groups. Only in relation to the objective learning trajectory the pattern of over- and underconfidence differed across groups, that is, the tendency of younger adults to remain underconfident and of older adults to remain overconfident from the second trial on persisted throughout almost all conditions.

2.3.4 Conclusion

The UWP effect has been examined exclusively in young populations (see, e.g., Koriat et al., 2002) and, hence, it remained an open issue if this effect could be replicated in older adults. As has been hypothesized by Scheck and Nelson (2005), part of the UWP effect might be due to the impact of anchoring on the formation of JOLs. First of all, our results seem to support the notion of a psychological anchor which determines JOLs at the first trial. The anchoring effect appeared to be very pronounced when no prior information was available, as it was the case for immediate JOLs: Both age groups rated the probability of recalling the same items at almost identical levels. Simultaneously, the difficulty of items was very different for both groups as young recalled up to four times more items than older participants (as seen in Experiment 2 with easy words after the first trial). We interpret this in favour of an anchoring mechanism which determined largely the location of JOLs – almost independent of item difficulty. Note that the higher accuracy of delayed JOLs is not contradictory to this view. When judgments are delayed, JOLs are, presumably, based on a retrieval attempt, which provides a mnemonic cue that might be utilized in forming a more accurate JOL (Nelson & Dunlosky, 1991). A similar mechanism leads to increasingly accurate judgments when participants are given more than one trial, that is, the monitoring process following the first trial reduces the bias toward the anchor to a considerable extent and results in an increase in absolute accuracy (Koriat et al., 2002; Nelson & Dunlosky, 1991).

Earlier studies have shown that repeated practice leads to an average JOL level which typically falls below the level of memory performance in the second trial and consequently

instantiates the UWP effect. In fact, in these earlier studies the UWP effect appeared to be very robust, at least against several experimental manipulations, and we therefore expected it to appear in older adults as well (for a summary, see Koriat et al., 2002). The most striking finding from both our experiments was the non-appearance of the UWP effect in older adults at all, while the results in the younger group replicated earlier findings (Scheck & Nelson, 2005). As both groups received the same stimulus material and the same procedure, it seemed unlikely that methodological differences were responsible for the unexpected results in the older group. The presented paired-associates, however, appeared to be much more difficult for older than for younger participants. Still, we concluded that older adults do not display the UWP effect as found in young people.

Several factors may be responsible for the non-occurrence of the effect: As mentioned earlier, throughout both experiments and across all learning trials younger adults remembered more items than older adults implying that for older adults the same items were more difficult. This is not an unusual finding in laboratory test situations when paired associates are presented in non-self paced learning trials. If older and younger adults are compared in their recall performance, young participants generally recall more words than older participants and, moreover, they tend to learn faster than older adults (Kausler, 1994). This was also observed in both our experiments where younger adults recalled more items and learned faster which resulted in steeper learning trajectories compared to older participants' performance. Even though the stimulus was the same for both age groups, the difficulty appeared to be highly elevated for older adults. The older group started at very low recall levels and did not benefit much from additional learning trials. Given these circumstances, the preconditions for the UWP effect were not exactly the same for young and old participants. Equal JOL levels across both groups but lower recall levels in the older lead to a larger overconfidence effect in the old group after the first trial. Further, young participants benefited more from additional learning trials than older, resulting in steeper learning trajectories. Note that the UWP effect requires an interaction between the average JOL and recall trajectories, implying JOLs in the second trial to be significantly underconfident, which, altogether, increases the demands with respect to the UWP effect for older adults: Older adults would have been required to adapt their JOLs in the second trial to a greater extent compared to younger adults in order to elicit underconfident judgments. In fact, older adults adjusted their JOLs, from the first to the second trial, to a greater extent than younger, but still not enough to fall below the recall level.

Why did older adults not adapt their JOLs to a greater extent? There might be two explanations: One is that JOLs can not be adapted arbitrarily away from a given level but only to a certain degree due to a persistence or inertia of JOLs. That is, if the average JOL in the first trial was at 40% one will not simply downgrade it by 35 points to 5% but, for example, maximally by 15 points to 25%. Hence, the adaptation of JOLs from the first to the second trial maybe tapped the full range in older adults adjustment possibilities. Even with the maximal downward adaptation of JOLs, this was not enough to elicit the UWP effect because recall was still lower. The second explanation bears on earlier findings where older adults tended to generally overestimate their cognitive performance (Connor et al., 1997; Murphy et al., 1981; Touron & Hertzog, 2004). If this was the case for JOLs in both of our studies, the instantiation of the UWP effect must have been additionally impeded simply because the required underconfident judgements are all biased with the general tendency to overestimate one's own performance and, finally, lead to higher average JOLs.

Instead of just focusing on the presence or absence of the UWP effect in older adults, our experiments have shown that it might be also fruitful to concentrate on (dis-)similarities in JOLs in young and old participants across all learning trials to learn more about monitoring. The JOL trajectories in older and younger adults were fairly similar in their shape: Young and old adults both showed the largest adaptation in JOLs from the first to the second trial, which was also underlined by significant interaction terms of Trial × Measure. From the second trial on, the adjustment of JOLs was much smaller but still lead to increasingly accurate judgments. By and large, the adaptation process in JOLs was comparable in young and old participants. Furthermore both groups started out, in the immediate condition, at almost exactly the same level which suggests that the anchor is not underlying much change in two very different cohorts. One might speculate if younger adults would have shown an UWP effect if their learning performance had been as low as in older adults, but the basic process of JOLs adjustment over a number of trials appeared to be basically the same in the young and the old group. On the other hand, the relation between JOLs and recall performance from trials two to five was very different between both groups. As the young underestimated their performance, older adults systematically overestimated their recall level, especially when JOLs were given immediately. These results are similar to findings from global memory predictions where older adults tend to overestimate their memory performance as well. Murphy, Sanders, Gabriesheski, and Schmitt (1981), for example, had younger and older

adults estimate their memory span for the number of common objects that they thought they could remember, followed by a recall task in which this span was actually measured. They found that younger adults tended to underestimate their memory span, whereas older adults tended to overestimate their memory span.

In view of the three explanatory approaches described in the introduction, the dual-factor hypothesis (Scheck et al., 2004) appeared to best catch the interplay between a strong anchor in the first trial and increasing importance of monitoring throughout the following trials. The higher accuracy for JOLs in the delayed condition can be deducted by the dual factors hypothesis as well: Delayed presentation of the cue word initiates a first retrieval attempt which delivers valuable information on the probability of recalling the item later. Consequently, monitoring outweighs the arbitrary anchor in the formation of the delayed JOL and leads to greater accuracy. Note, however, that in all three approaches, in the cue-utilization framework, in the anchoring hypothesis, and in the dual-factors hypothesis, the weighing of different cues is crucial for the outcome of a response. Hence, if one considers the anchor to represent an internal cue which may be used in absence of any other relevant cue of item difficulty, the anchoring hypothesis and the dual-task hypothesis may be recast in terms of the cue-utilization approach.

In the introduction to this paper we emphasized the importance of monitoring memory, especially in old age, and argued that monitoring is spared from cognitive decline. In fact, the process of memory monitoring did not seem to be different from monitoring in the younger group except for its tendency to be overconfident. This, however, puts into perspective the advantage of an intact monitoring in older persons. If one overestimates his own memory performance, the effort invested in future learning trials is probably smaller compared to someone who underestimates his or her performance. Unfortunately, the older group with low recall performance and small learning rates overestimated their performance and, maybe, this contributed to even lower recall levels. If monitoring is to be used in future, as an intact resource for memory enhancement, older adults may benefit from it when their judgments become underconfident. Hence, the presence of the UWP effect could be seen as an indicator of an intact and self-propelling memory system.

3 General discussion

The aim of the present thesis was to find new theoretical and empirical accounts for the low correspondence between memory self-reports and memory performance. More specifically, the present work adressed three research questions regarding (1) the invariance in the measurement of memory self-reports, (2) the relation between self-reports and learning, and (3) the relation between monitoring and learning. The research on these three questions was motivated by the role metamemory could play in the context of aging and diminishing cognitive resources. It is well documented that, on average, memory performance declines into old age (cf. Kausler, 1994; Rönnlund et al., 2005) but there is also evidence that decline can be compensated at a biological (Grady & Craik, 2000; McIntosh et al., 1999) as well as at the psychological level (Cavanaugh & Blanchard-Fields, 2006). At the psychological level, metamemory could represent the key to both, a successful self-diagnosis of the general memory ability and, consequently as the basis for an optimal allocation of remaining cognitive resources to memory processes. The scientific investigation of the interplay between memory self-reports and memory performance, however, did not support the expectations placed on metamemory. The strongest argument against the designated role metamemory could assume is the low accuracy of its reports. That is, metamemory judgments are typically found to correlate only moderately with their respective behavioral counterpart, namely, memory performance (see Chapter 1.3). In that case it is questionable if metamemory can effectively compensate memory decline because only a realistic estimation of ones' own memory entails that people are able to adopt appropriate memory strategies. However, it remains uncertain if the low accuracy in metamemory judgements reflects a true incapacity to introspect about one's own memory or if it is rather due to inadequate operationalization of memory self-reports or memory performance.

As a consequence of the low relation between self-report and memory performance, research on metamemory has turned away from the initial straightforward hypothesis that memory self-reports actually reflects memory performance (Kail, 1990; Schneider, 1985) and has moved its focus to the investigation of the low correspondence between subjective and objective memory measures. The focus in this thesis is also on possible reasons for the low relation between subjective and objective memory, however, I explicitly intended to revaluate the straightforward hypothesis by examining three open research questions. The first research

question addressed possible sources of bias in the memory self-report measures and the second in the operationalization of memory. The third question pertained to age differences in monitoring in a multitrial (i.e., learning) setting and, hence, it addressed the effect of aging on the accuracy of monitoring.

3.1 Summary and discussion of the results

3.1.1 Measurement invariance of the cognitive failures questionnaire across the adult lifespan

The first research question pertained to whether the measures used to operationalize memory self-reports are invariant across age, and thus measure the same constructs in each age group. In order to test for measurement invariance (MI) the commonly used Cognitive Failures Questionnaire (CFQ; Broadbent et al., 1982) was chosen because it not only taps memory failures but also other areas of cognitive failures which might be less prone to implicit theories about aging (Cavanaugh et al., 1998). Invariance hypotheses were systematically tested across six age groups ranging in age from 24 to 83 years. Although the CFQ is a widely used instrument, its factor structure remained an issue of scientific debate. The study reported in Chapter 2.1 used data of a representative sample (N = 1,303) from the Maastricht Aging Study (MAAS) to test and compare factor solutions for the CFQ previously reported in the literature by means of confirmatory factor analysis. Additionally, a three-factor model of the CFQ that emerged from an exploratory factor analysis was examined. In order to minimize biased parameter estimates from treating Likert-type scale items as continuous, the factor analyses and the consecutive investigation of MI was based on ordered categorical variables (DiStefano, 2002; Lubke & Muthén, 2004; O'Brien, 1985). The three-factor model was tested for increasing levels of measurement invariance across six age groups. The CFQ proved to be measurement invariant across all age groups and, hence, the mean trajectories as well as the variances and covariances of the three factors were meaningfully interpretable across age. Even though all three factors represent cognitive domains, they showed very different trajectories across age groups. The mean of the memory factor termed *Forgetfulness* increased throughout the six agegroups almost linearly, that is, older adults reported more forgetfulness compared to younger adults. This is a classic finding which corroborates results from earlier studies (cf. Kausler, 1994). The factor measuring *Distractibility*, in turn, followed a different trajectory. Namely, the two oldest groups reported significantly less distractibility

than the four younger groups. For the third factor, *False Triggering,* no age-related increase or decline resulted in the means. Interestingly, factor (co-)variances remained stable across the age groups. This indicates that the relation between the three factors does not depend on the age of the respondent. The same can be concluded for the variability of the three factors. It further implies that, even if contextual influences including history-graded and non-normative influences (Baltes, Reese, & Lipsitt, 1980) affect cognition, its effects do not influence the measurement properties of the measures, that is, the constructs' measurement functions equivalently for each age group. This finding, however, does not exclude that contextual influences may affect mean-levels of memory self-reports or other types of change and continuity.

The present findings regarding MI of the CFQ explicitly demonstrated what was implicitly assumed in earlier studies – but not being tested systematically – namely, that strict MI would hold across age. That is, factor loadings, intercepts and residual variances in one group are equal to corresponding loadings, intercepts and residuals in other groups and, hence, factor mean differences are unambiguously interpretable. Even though the CFQ proved to be strictly measurement invariant, this can not be taken as a grant for other self-report questionnaires of cognitive failures or for the CFQ in other samples. MI depends largely on the quality of items purported to measure a certain latent construct (Lubke et al., 2003). Hence, different scales may show different degrees of MI which affects interpretation of, for example, age differences in another questionnaire so that the analysis of MI across a given selection variable has to be newly tested. In the case of the CFQ one might conclude that age differences are meaningfully interpretable and that differences in factor means are not artificial, but represent truly experienced change in the underlying factors (e.g., Forgetfulness). With the examination of MI, one has a tool at hand which allows controlling for bias in the case where groups are compared by means of questionnaire data. In this respect, the first research question might never be fully answered but rather represents an issue which is always present when different groups are compared. Hence, the procedure of assessing MI of a given measure should be part of the normal scientific workflow.

3.1.2 Verbal learning as an alternative memory measure

The second research question addressed the relation between memory self-reports and memory performance with the focus on learning as, possibly, a more informative memory

measure. The results presented herein bear only indirectly on the issue of low correspondence between memory self-reports and memory measures as they were mainly used to demonstrate the methodological approach to formalize and relate learning to other variables of interest. However, they suggest one line of research that might be pursued to scratch the surface of the intricate and long-standing issue of the correspondence between subjective and objective memory reports. In Chapter 1.4.2, I argued that data from learning experiments would deliver a closer relation to memory self-reports elicited in memory questionnaires. Hence, other than the first research question, here, the focus was on the objective memory measure. There were two reasons for focusing on the memory measure: First, the approach to modify self-report questionnaires with respect to behavioral specificity (cf. Hertzog et al., 2000) did not seem to be a very compelling alternative because the increase in the correlation between subjective and objective memory measure was only marginal and, at the same time, it was outweighed by the loss in generalizability with respect to the subjective measure. Second, in expanding the memory measure from a single trial recall task to a learning experiment with several study and test trials, yielded a more broad and differentiated view on memory. Furthermore, in Chapter 1.4.2, I argued that memory is probably apprehend by lay persons in a more generalized way, that is, in a naturalistic situation memory is not reduced to a single recall event but the whole process of information acquisition and recall is considered to represent memory. Consequently, if subjects are asked to give reports about their memory performance the behavioral correlate which is measured in a memory test must coincide to the largest possible amount with the subjective measure. In this respect, learning experiments seem to be better suited than one-trial memory tests.

The approach presented in Chapters 1.4.2.1 and 1.4.2.2 was demonstrated by means of data from 364 persons from the Zurich Longitudinal Study on Cognitive Aging (ZULU; Zimprich et al., in revision). Part of the testing protocol was the Metamemory in Adulthood (MIA) questionnaire, three measures assessing processing speed, and a verbal learning measure that comprised five study and recall cycles. The study was mainly geared to demonstrate the fruitfulness of the individually-centered approach on verbal learning in old age. In a first step the best representation of verbal learning was assessed by testing three different nonlinear functions, namely, the quadratic, the exponential, and the hyperbolic growth curve. The latter turned out to fit the data best and the linkages of the learning parameters initial performance (β), potential maximum performance (α), and rate of learning

(γ) to age and processing speed were modelled. The inclusion of the latter cognitive ability was motivated by the fact that processing speed represents a major explanatory variable of cognitive aging (Salthouse, 1991, 1996). Subsequently, we included memory as an outcome variable, that is, verbal learning parameters (β, α, γ) were used as predictor variables of memory performance in old age. In light of the duality of learning and memory phenomena, memory performance represents an obvious and, moreover, extensively studied outcome variable in cognitive aging research (Craik, 1977; Hultsch et al., 1998; Kausler, 1994). In the complete model, the verbal learning parameters initial performance level (β), potential maximum performance (α), and rate of learning (γ) thus acted as mediating variables between processing speed and memory.

The variations in the three learning parameters indicating individual differences were highest in the learning rate (γ) which suggests that older persons tend to show more pronounced individual differences from each other in the rate of acquisition than in initial performance or potential maximum performance. Further, the learning rate seemed to be mostly affected by age, whereas initial performance was less and potential maximum performance was almost not affect by age. That is, older participants needed more trials to reach their maximum performance, which, in turn, was not affected by the participants' age. This supports the idea formulated in Chapter 1.4.2 where I argued that the recall performance might remain stable, but the cognitive effort to maintain a given level increases in older adults. The influence of age, however, was in part mediated by processing speed, which is in line with Salthouse's (1996) processing speed theory. Eventually, memory, measured in single trial recall tests, was added as an outcome variable of the three learning parameters. Interestingly the learning rate and the potential maximum performance exerted the largest effect on the memory measures, which implies that these two parameters are determining and inherent factors in memory.

In sum, the study presented in Chapter 2.2 demonstrated the benefit one gains from administering several learning and recall trials over the commonly used single trial experiments. That is, at least two more parameters related to memory performance are estimable which appear to be memory-inherent. Further, due to increasing individual differences the parameters can be mapped more exactly and unsystematic influences on performance become smaller which should increase reliability of the measure. The question whether the relatedness between self-referent memory beliefs and learning parameters are

higher compared to single trial memory measures, however, has not directly been tested and can not be answered at this point.

3.1.3 Monitoring and learning in young and old adults

The third research question concerned age differences in the relation between monitoring and learning by means of a cross-sectional design. Monitoring might be considered as a process which functions on-line, that is, which delivers information to the meta-level about the current state a monitored object is in. On the basis of the information delivered by monitoring, controlling processes may take place to alter the state of the object, which, again, will be subject to monitoring. The importance of such monitoring processes may gain even more weight when memory performance is declining, as it is the case with older adults. One way to operationalize monitoring is by eliciting Judgments-of-Learning (JOLs) (see Chap 1.2.1 or refer to Nelson & Narens, 1994). Apart from the finding that JOLs are only moderately accurate, Koriat (1997) reported a discrepancy between JOLs and recall performance which he termed underconfidence-with-practice (UWP) effect (see Chapter 1.4.3): As recall performance increased from the first to the second trial, the effects of practice did not lead to more accurate predictions. Instead of improving calibration, average JOLs for the second occasion became markedly lower (underconfident) than recall performance.

An open issue is whether the UWP effect could also be found in older adults. To that end, two experiments were conducted in order to shed light on age differences in JOLs and more specifically to estimate the susceptibility in older adults regarding the UWP effect. In the first experiment, both younger and older adults overestimated their memory performance in a first trial, in fact, the JOLs seemed to be determined largely by an anchoring effect. In the second trial, the older group differed from the young group. The JOLs given by the younger participants underestimated significantly the recall performance whereas JOLs given by older participants matched, on average, the correctly recalled words. Actually, the UWP effect was not observed in any of several conditions in older participants. The complete absence of the UWP effect in the older age group seemed even more remarkable if one considers that the effect has been shown to withstand several manipulations and, thus, was thought to be very robust (cf. Koriat et al., 2002). In the second experiment involving five study-test cycles, the same basic pattern of results was present: Young and old participants' first JOLs appeared to

be largely biased by an anchor. In the consecutive trials younger showed the UWP effect but older adults still overestimated or predicted correctly their recall performance. The findings appeared to fit into a framework of dual factors affecting JOLs, which posits that the magnitude of JOLs derives both from an anchoring point and from on-line monitoring of the items.

In respect of the research question formulated in Chapter 1.4.3, which pertained to age differences in monitoring, the findings indicate both similarities and differences between age groups. That is, if no prior information about item difficulty was available older and younger adults started off at almost identical JOL levels. This is remarkable given that the same stimuli were significantly more difficult for older than for younger adults. Hence, JOLs provided in the first trial appeared to be independent from true item-difficulty and the effect of the anchor was almost identical in both age groups. With every additional trial, however, the discrepancy between JOLs and recall performance diminished and, overall, accuracy increased at the same time. Note that this finding also underlines the need for multitrial memory tasks when monitoring is examined because monitoring accuracy obviously is not a static measure, but changes significantly over time. A further similarity between young and old participants concerned the shape of the mean JOLs curve: At first, participants largely overestimated their recall performance. After the initial overconfidence, the average JOLs in the second trial were corrected largely toward a lower mean. In the consecutive trials the form of the mean JOLs curve was increasing and the correction in JOLs from one trial to the other were smaller compared to the adjustment between the first and the second trial (see, for example, Figure 2.6).

The main difference between both age groups became obvious only when JOLs were contrasted to the actual recall performance. That is, young adults displayed the UWP effect and approximated the recall performance from underconfident judgments. In turn, older adults were, on average, never underconfident and, hence, did not display the UWP effect, and approximated the performance in recall from overconfident judgments. Note that in terms of absolute accuracy both groups were similar, in fact, older adults tended to predict their recall performance on average more accurately than younger. This might highlight a crucial difference in older adults monitoring functioning: If older adults judge their recall performance too optimistically, or correctly, a consequence might be that they invest less cognitive resources in memory-related tasks and, hence, their memory performance is lower

as it could be. Hence, one might speculate that a certain amount of underconfidence keeps up the learning effort and has a positive effect on memory performance.

3.2 Coda and outlook

The unifying element in the three research questions was the issue of correspondence between memory self-reports and memory performance. The questions aimed at three neglected topics in metamemory research ranging from methodological considerations about the application of memory questionnaires, over an alternative operationalization of memory, to age differences in monitoring. Admittedly, the range of these questions was very broad as they covered different methodological issues, but also with regard to content the metacognitive functions subsumed under the term "memory self-reports" turned out to be very heterogeneous. That is, self-referent memory beliefs are probably part of a persons' self-concept (Silvia & Gendolla, 2001), hence, they are rather stable across the adult lifespan and are retrievable from memory upon request. JOLs, in turn, are part of monitoring which is best considered as a process and, hence, JOLs can differ considerably in a very short period of time. Still, the initial question about the low relation between subjective and objective memory measures is so ubiquitous in metamemory research that it concerns almost all studies in this field. Hence, the present thesis did not intend to find conclusive answers to the open research questions.

The first two research questions pertaining to MI and the relation between self-reports and learning remained open but they might be incorporated in future research. That is, by determining the degree of MI the risk of misinterpreting group differences can be minimized to a manageable degree. This does not necessarily imply that results from older studies need to be rewritten. For example, the investigation of the CFQ yielded a strictly measurement invariant scale which also implies that inferences drawn with that instrument, regarding age differences are not biased by differential item functioning. Note, however, that MI is tied to a specific factor structure and, as a consequence, the same set of items in a different factor solution can not be measurement invariant because MI relies on the assumption that the model is "true". That is, with regard to the CFQ, earlier solutions, for example the four factors solution presented by Wallace and colleagues (2002), might not be MI and conclusions regarding group differences drawn upon this basis of factors might not be meaningfully interpretable. This is a limitation which mainly concerns questionnaires with debatable factor

solutions (e.g., the CFQ). However, it does not necessarily apply for other questionnaires/scales as well. The MIA, for example, is composed of seven subscales which, simultaneously, represent the factor structure of the measure. Unlike in the case of the CFQ, the factor structure of the MIA is theoretically derived which also implies two things: First, if the MIA proves to be measurement invariant across groups, this probably holds for other similar samples as well and research based on that questionnaire might be corroborated ex post. Second, a confirmatory factor solution based on a theoretical instead of an empirical classification of factors may be more difficult to verify than an item solution which emerged from an exploratory factor solution which is gauged on a specific sample. This, however, may also imply that self-report questionnaires, such as the MIA, are more prone to differential item functioning because items may not load high enough on the respective factors and, hence, high degrees of MI can not be achieved.

Similar to the first, also the second research question remains an open issue. By introducing learning as a broader memory measure the focus is shifted away from the self-reports to the objective measure. The main aim of the present thesis was to find additional accounts for the low relation between subjective memory reports and the respective memory performance. Previous research on this low relation has mainly focused on the subjective measure; in contrast, the measures of memory performance did not receive much attention. This neglected topic should be dealt with in the future in order to gain a broader understanding of the relation between subjective and objective memory performance. Therefore in Chapter 1.4.2 an approach was presented which might give additional accounts of the interplay between self-reports and memory performance. By identifying additional memory inherent parameters, apart form the recall performance after one trial, the process of learning can be recast in its relevant elements, that is, initial performance, learning rate, and potential maximum performance. As shown in the second study, these parameters proved to be differentially affected by age-related decline in processing speed. Furthermore, these parameters appear to also have differential impact on recall performance in other memory tests. This offers a much more differentiated view on memory functioning and the strength in age-related change in memory performance can be identified for each parameter separately. If different memory parameters are affected by decline in cognitive resources unequally this also has implications for other areas in metamemory research. As pointed out in Chapter 1.2.1 accurate self-reports could be used as self-diagnosis for memory complaints. But, as shown in

a number of studies (Hänninen et al., 1994; Niederehe & Yoder, 1989; Zimprich & Kliegel, in press) memory complaints do not correlate highly with memory performance. It would be premature, however, to conclude that people are generally not good at introspecting and reporting their memory. Maybe, experienced change in memory performance is better captured by the learning rate, instead of the initial performance. And if people report about their memory capacities they might refer to potential maximum performance, which is not captured by a single-trial memory task. Further, note that, for example, the correlation between potential maximum performance and initial memory performance is r = .29 (cf. Figure 2.3), which is similarly low compared to the correlations found between memory self-reports and the recall performance in single-trial memory tasks. Hence, correlating self-reports with initial memory performance may lead to low values because the true correspondence is between the self-report and potential maximum performance. In the future, it might prove worthwhile to investigate and clarify the relations between different types of memory self-reports and memory parameters. A further implication drawn from these results concerns the aspect of memory improvement or compensation with declining cognitive resources. Instead of improving memory in general, it might be more fruitful to focus on the most relevant memory parameters. This has the advantage that cognitive resources are invested where they have the greatest impact. For example, improving the learning rate has a greater effect on memory performance than improving recall performance after one trial. Hence, a deeper and more differentiated understanding of memory self-reports and memory performance is not only of scientific interest but could also avail other domains as, for example, mnemonic training. In summarizing, future research on metamemory should be open to new developments in methodology to measure micro- and macrodevelopment in memory and self-reports by utilizing latent curve models which allow modelling both fixed effects and individual departures from these average effects (Blozis, 2004; Browne, 1993).

The third topic which remained neglected in the scientific investigation of metamemory was age-related differences in memory monitoring. Other than self-referent memory beliefs, monitoring is a process which delivers on-line information about a current memory-state and, hence, it can have a direct and immediate influence on memory performance (see Chapter 1.2.1). Hence, monitoring is better captured with multi-trial experiments which regard for changes in the monitoring process. The third study presented in this thesis revealed differences in monitoring, mainly in the pattern of over- and

underconfidence. Most importantly, older adults appeared to overestimate the probability of recalling a given set of items, not only in the first, but also in the consecutive learning trials. Younger adults, in turn, displayed the UWP effect and, hence, underrated their recall performance from the second trial on. Note that both experiments bridged an important gap: The UWP effect was thought to be very robust and omnipresent (Koriat et al., 2002), instead, older adults did not display the effect at all. It appeared that the formation of JOLs did not change markedly across the adult lifespan, that is, the first monitoring attempt is guided by an anchoring process which represents the best guess about the probability of later recalling successfully a given item. With every additional trial, however, the accuracy of monitoring increased until a very close correspondence was achieved. This implies that people are able to monitor correctly their memory performance if sufficient trials are provided. Apparently, the main difference between older and younger adults is not the accuracy, but the pattern of over- and underconfidence. This difference is captured by the UWP effect which, as argued in Chapter 2.3.4, may be an indicator of a well functioning and self-propelling memory system: The effort to invest cognitive resources is best kept up if the goal feels always a little out of reach. Older adults appear to be too confident about their memory performance so that they have no reason to intensify their cognitive effort. This assumption, however, needs to be tested in future studies investigating the effect of being over- or underconfident on memory performance.

In sum, apart from the three discussed research questions, the correspondence between metamemory and memory performance remains to large portions a research field with a number of uncertainties, especially in the lifespan view. Apart from highlighting three open research questions this thesis also demonstrated that metamemory still has the potential to be the most important psychological mechanism to compensate age-related memory decline.

References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117*, 288-318.

Albert, S. (1977). Temporal comparison theory. *Psychological Review, 84*, 485-503.

Allemand, M., Zimprich, D., & Hertzog, C. (2007). Cross-sectional age differences and longitudinal age changes of personality in middle adulthood and old age. *Journal of Personality Assessment, 75*, 323-358.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought.* Mahwah, NJ: Erlbaum.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*, 396-408.

Antonini, A., Leenders, K. L., Reist, H., Thomann, R., Beer, H. F., & Locher, J. (1993). Effect of age on d2 dopamine receptors in normal human brain measured by positron emission tomography and 11c-raclopride. *Archives of Neurology, 50*, 474-480.

Arbuckle, T. Y., Gold, D., & Andres, D. (1986). Cognitive functioning of older people. *Psychology and Aging, 1*, 55-62.

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling: A Multidsiciplinary Journal, 12*, 411-434.

Bäckman, L., & Farde, L. (2004). The role of dopamine systems in cognitive aging. In R. Cabeza, L. Nyberg & D. C. Park (Eds.), *Cognitive neuroscience of aging: Linking cognitive and cerebral aging* (pp. 58-84). New York: Oxford University Press.

Baddeley, A. (1990). *Human memory: Theory and practice* (Revised ed.). Needham Heights: Allyn & Bacon.

Baltes, P. B., & Kliegl, R. (1992). Further testing of limits of cognitive plasticity: Negative age differences in a mnemonic skill are robust. *Developmental Psychology, 28*, 121-125.

Baltes, P. B., & Lindenberger, U. (1997). Emergence of a powerful connection between sensory and cognitive functions across the adult life span: A new window to the study of cognitive aging? *Psychology and Aging, 12*, 12-21.

Baltes, P. B., Reese, H. W., & Lipsitt, L. P. (1980). Lifespan developmental psychology. *Annual Review of Psychology, 31*, 65-110.

Barker, A., Carter, C., & Jones, J. (1994). Memory performance, self-reported memory loss and depressive symptoms in attenders at gp-referal and a self-referal memory clinic. *International Journal of Geriatric Psychiatry, 9*, 305-311.

Bauer, D. J. (2005). The role of nonlinear factor-to-indicator relationships in tests of measurement equivalence. *Psychological Methods, 10*, 305-316.

Bedard, M., Leonard, E., McAuliffe, J., Weaver, B., Gibbons, C., & Dubois, S. (2006). Visual attention and older drivers: The contribution of inhibition of return to safe driving. *Experimental Aging Research, 32*, 119-135.

Biesanz, J. C., Deeb-Sossa, N., Papadaki, A. A., Bollen, K. A., & Curran, P. J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods, 9*, 30-52.

Bloem, R., & Schmuck, P. (1999). Individual differences in cognitive inhibition and their relation to failures of attention. *Diagnostica, 45*, 47-55.

Blozis, S. A. (2004). Structured latent curve models for the study of change in multivariate repeated measures. *Psychological Methods, 9*, 334-353.

Bolla, K. I., Lindgren, K.-N., Bonaccorsy, C., & Bleecker, M. L. (1991). Memory complaints in older adults: Fact or fiction? *Archives of Neurology, 48*, 61-64.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Boomsma, D. I. (1998). Genetic analysis of cognitive failures (cfq); a study of dutch adolescent twins and their parents. *European Journal of Personality, 12*, 321-330.

Broadbent, D. E., Cooper, P. F., FitzGerald, P., & Parkes, K. R. (1982). The cognitive failures questionnaire (cfq) and its correlates. *British Journal of Clinical Psychology, 21*, 1-16.

Brown, R., & Middendorf, J. (1996). The underestimated role of temporal comparison: A test of the life-span model. *Journal of Social Psychology, 136*, 325-331.

Brown, S., & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behavior Research Methods, Instruments, and Computers, 35*, 11-21.

Browne, M. W. (1993). Structured latent curve models. In C. M. Cuadras & C. R. Rao (Eds.), *Multivariate analysis: Future directions 2* (pp. 171-197). Amsterdam: Elsevier Science.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Browne, M. W., & Du Toit, S. H. C. (1991). Models for learning data. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 47-68). Washington, DC: American Psychological Association.

Bruce, E. R., Coyne, A. C., & Botwinick, J. (1982). Adult age differences in metamemory. *Journal of Gerontology, 37*, 354-357.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning*. New York: Wiley.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.

Cavanaugh, J. C., & Blanchard-Fields, F. (Eds.). (2006). *Adult development and aging* (5th ed.). Belmont, CA: Wadsworth Publishing/Thomson Learning.

Cavanaugh, J. C., Feldman, J. M., & Hertzog, C. (1998). Memory beliefs as social cognition: A reconceptualization of what memory questionnaires assess. *Review of General Psychology, 2*, 48-65.

Cerella, J., Onyper, S. V., & Hoyer, W. J. (2006). The associative-memory basis of cognitive skill learning: Adult age differences. *Psychology and Aging, 21*, 483-498.

Chapman, G. B., & Johnson, E. J. (1999). Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes, 79*, 115-153.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling: A Multidsiciplinary Journal, 12*, 471-492.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 233-255.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 112*, 558-577.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metameory accuracy. *Psychology and Aging, 12*, 50-71.

Cousineau, D., Hélie, S., & Lefebvre, C. (2003). Testing curvatures of learning functions on individual trial and block average data. *Behavior Research Methods, Instruments, and Computers, 35*, 493-503.

Coyne, A. C. (1985). Adult age, presentation time, and memory performance. *Experimental Aging Research, 11*, 147-149.

Craik, F. I. M. (1977). Age differences in human memory. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 384-420). New York: Von Nostrand Reinhold.

Craik, F. I. M., Anderson, N. D., Kerr, S. A., & Li, K. Z. H. (1995). Memory changes in normal ageing. In A. D. Baddely, B. A. Wilson & F. N. Watts (Eds.), *Handbook of memory disorders* (pp. 211-241). New York: Wiley.

Crook, T. H. I., & Larrabee, G. j. (1990). A self-rating scale for evaluating memory in everyday life. *Psychology and Aging, 5*, 48-57.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105*, 317-327.

Cudeck, R., & Harring, J. R. (2007). The analysis of nonlinear patterns of change with random coefficient models. *Annual Review of Psychology, 58*, 615-637.

Davidian, M., & Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data*. New York: Chapman & Hall.

Davis, H. P., Small, S. A., Stern, Y., Mayeux, R., Feldstein, S. N., & Keller, F. R. (2003). Acquisition, recall, and forgetting of verbal information in long-term memory by young, middle-aged and elderly individuals. *Cortex, 39*, 1063-1091.

Dellenbach, M., & Zimprich, D. (in press). Typical intellectual engagement and cognition in old age. *Aging, Neuropsychology, and Cognition*.

Derouesné, C., Lacomblez, L., Thibault, S., & LePoncin, M. (1999). Memory complaints in young and elderly subjects. *International Journal of Geriatric Psychiatry, 14*, 291-301.

Devolder, E. A., Brigham, M. C., & Pressley, M. (1990). Memory performance awareness in younger and older adults. *Psychology and Aging, 5*, 291-303.

Devolder, P. A., & Pressley, M. (1991). Memory complaints in younger and older adults. *Applied Cognitive Psychology, 5*, 443-454.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 327-346.

Dixon, R. A., Hultsch, D. F., & Hertzog, C. (1988). The metamemory in adulthood (mia) questionnaire. *Psychopharmacology Bulletin, 24*, 671-688.

Dunlop, K. (1912). The case against introspection. *Psychological Review, 19*, 404-413.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (jols) and the delayed-jol effect. *Memory & Cognition, 20*, 373-380.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (jols) to the effects of various study activities depend on when the jols occur? *Journal of Memory and Language, 33*, 545 -565.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*, 215-251.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review, 57*, 94-107.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*, 117-140.

Finney, S. J., & DiStefano, C. (2006). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269-314). Greenwich, CT: Information Age Publishing.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*, 286-299.

Gall, T. L., Evans, D. R., & Howard, J. (1997). The retirement process: Changes in the well-being of male retirees across time. *Journal of Gerontology: Psychological Sciences, 52B*, 110-117.

García Martínez, J., & Sánchez-Cánovas, J. (1994). Adaptation of cognitive failures questionnaire by broadbent, cooper, fitzgerald & parkes. *Analisis y Modificacion de Conducta, 20*, 727-757.

Gilewski, M. J., Zelinski, E. M., & Schaie, K. W. (1990). The memory functioning questionnaire for assessment of memory complaints in adulthood and old age. *Psychology and Aging, 5*, 482-490.

Goldstein, H. (Ed.). (1995). *Multilevel statistical models* (2 ed.). London: Arnold.

Goldstein, H., & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, 158A*, 175-177.

Grady, C. L., & Craik, F. I. M. (2000). Changes in memory processing with age. *Current Opinion in Neurobiology, 10*, 224-231.

Greenfield, J. P., Blackwood, W., McMenemy, W., Meyer, A., Norman, R., & Russel, D. (1967). *Neuropathology*. Baltimore: Williams & Wilkins.

Hänninen, T., Reinikainen, K. J., Helkala, E. L., Koivisto, K., Mykkänen, L., Laakso, M., et al. (1994). Subjective memory complaints and personality traits in normal elderly subjects. *Journal of the American Geriatrics Society, 42*, 1-4.

Härting, C., Markowitsch, H. J., Neufeld, H., Calabrese, P., Deisinger, K., & Kessler, J. (2000). *Wechsler gedächtnistest - revidierte fassung*. Bern: Huber.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, 356-388.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin and Review*, 185-207.

Heckhausen, J., Dixon, R. A., & Baltes, P. B. (1989). Gains and losses in development throughout adulthood as perceived by different adult age groups. *Developmental Psychology, 25*, 109-121.

Helmstädter, C., Lendt, M., & Lux, S. (2001). *Verbaler lern- und merkfähigkeitstest*. Göttingen: Lelitz.

Herrmann, D. J. (1982). Know thy memory: The use of questionnaires to assess and study memory. *Psychological Bulletin, 92*, 434-452.

Hertzog, C., & Dixon, R. A. (2005). Metacognition in midlife. In S. L. Willis & M. Martin (Eds.), *Middle adulthood: A lifespan perspective* (pp. 355-379). Thousand Oaks, CA: Sage Publications, Inc.

Hertzog, C., & Hultsch, D. F. (2000). Metacognition in adulthood and old age. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (2 ed., pp. 417-466). London: Lawrence Erlbaum.

Hertzog, C., Hultsch, D. F., & Dixon, R. A. (1998). Evidence for the convergent validity of two self-report metamemory questionnaires. *Developmental Psychology, 25*, 687-700.

Hertzog, C., Kidder, D. P., Powell-Moman, A., & Dunlosky, J. (2002). Aging and monitoring associative learning: Is monitoring accuracy spared or impaired? *Psychology and Aging, 17*, 209-225.

Hertzog, C., Park, D. C., Morrell, R. W., & Martin, M. (2000). Ask and ye shall receive: Behavioural specificity in the accuracy of subjective memory complaints. *Applied Cognitive Psychology, 14*, 257-275.

Hofer, S. M., & Sliwinski, M. J. (2001). Understanding ageing: An evaluation of research designs for assessing the interdependence of ageing-related changes. *Gerontology, 47*, 341-352.

Hofer, S. M., & Sliwinski, M. J. (2006). Design and analysis of longitudinal studies on aging. In K. W. Schaie & J. E. Birren (Eds.), *Handbook of the psychology of aging* (6 ed., pp. 15-37). San Diego, CA: Academic Press.

Horn, J. L., & Cattell, R. B. (1966). Age differences in primary mental ability factors. *Journal of Gerontology, 21*, 210-220.

Horn, J. L., & Hofer, S. M. (1992). Major abilities and development in the adult period. In R. J. Sternberg & C. A. Berg (Eds.), *Intellectual development* (pp. 44-99). New York: Cambridge University Press.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 117-144.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structrue analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Hultsch, D. F., Hertzog, C., Dixon, R. A., & Small, B. J. (1998). *Memory change in the aged*. New York: Cambridge University Press.

Ilmberger, J. (1988). *Münchner verbaler gedächtnistest*. München: Institut für Medizinische Psychologie.

James, W. (1890). *The principles of psychology* (Vol. 1). New York: Dover.

Jolles, J., Houx, P. J., van Boxtel, M. P. J., & Ponds, R. W. H. M. (Eds.). (1995). *Maastricht aging study: Determinants of cognitive aging*. Maastricht, the Netherlands: Neuropsych Publishers.

Jones, G. V., & Martin, M. (2003). Individual differences in failing to save everyday computing work. *Applied Cognitive Psychology, 17*, 861-868.

Kail, R. V. (1990). *The developement of memory in children* (3rd ed.). New York: Freeman.

Kausler, D. H. (1994). *Learning and memory in normal aging*. San Diego: Academic Press, Inc.

Kliegel, M., & Zimprich, D. (2005). Predictors of cognitive complaints in older adults: A mixture regression approach. *European Journal of Ageing, 2*, 13-23.

Klumb, P. L. (1995). Cognitive failures and performance differences: Validation studies of a german version of the cognitive failures questionnaire. *Ergonomics, 38*, 1456-1467.

Klumb, P. L. (2001). Tying knots in handkerchiefs: The use of memory aids in everyday life. *Zeitschrift fur Entwicklungspsychologie und Padagogische Psychologie, 33*, 42-49.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349-370.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*, 36-69.

Koriat, A., Ma'ayan, H., Sheffer, L., & Bjork, R. A. (2006). Exploring a mnemonic debiasing account of the underconfidence-with-practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 595-608.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*, 147-162.

Larson, G. E., Alderton, D. L., Neideffer, M., & Underhill, E. (1997). Further evidence on dimensionality and correlates of the cognitive failures questionnaire. *British Journal of Psychology, 88*, 29-38.

Larson, G. E., & Merritt, C. R. (1991). Can accidents be predicted? An empirical test of the cognitive failures questionnaire. *Applied Psychology: An International Review, 40*, 37-45.

Lindenberger, U., & Baltes, P. B. (1995). Kognitive leistungsfähigkeit im hohen alter: Erste ergebnisse aus der berliner altersstudie. *Zeitschrift für psychologie, 203*, 283-317.

Lineweaver, T. T., & Hertzog, C. (1998). Adults' efficacy and control beliefs regarding memory and aging: Separating general from personal beliefs. *Aging, Neuropsychology, and Cognition, 5*, 264-296.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492-527.

Logan, G. D. (1995). The weibull distribution, the power law, and the instance theory of automaticity. *Psychological Review, 102*, 751-756.

Lövdén, M., Ghisletta, P., & Lindenberger, U. (2004). Cognition in the berlin aging study (base): The first ten years. *Aging, Neuropsychology, and Cognition, 11*, 104-133.

Lovelace, E. A. (1990). Aging and metacognition concerning memory function. In E. A. Lovelace (Ed.), *Aging and cognition: Mental processes, self-awareness and interventions* (pp. 157-187). Amsterdam: North-Holland.

Lovelace, E. A., & Marsh, G. (1985). Predictions and evaluation of memory performance by young and old adults. *Journal of Gerontology, 40*, 197.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence, 31*, 543-566.

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to likert scale data complicates meaningful group comparisons. *Strucutral Equation Modeling, 11*, 514-534.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149.

Marsh, H. W., Balla, J. R., & Hau, K.-T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical properties. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling. Issues and techniques* (pp. 315-353). Mahwah, NJ: Lawrence Erlbaum.

Martin, M., Grünendahl, M., & Martin, P. (2001). Age differences in stress, social resources and well-being in middle and older age. *Journal of Gerontology: Psychological Sciences, 56*, 214-222.

Martin, M., & Zimprich, D. (2005). Cognitive development in midlife. In S. L. Willis & M. Martin (Eds.), *Middle adulthood: A lifespan perspective* (pp. 179-206). Thousand Oaks, CA: Sage Publications, Inc.

Matthews, G., Coyle, K., & Kraig, A. (1990). Multiple factors of cognitive failure and their relationship with stress vulnerability. *Journal of Psychopathology and Behavioral Assessment, 12*, 49-65.

Mazur, J. E., & Hastie, R. (1978). Learning as accumulation: A reexamination of the learning curve. *Psychological Bulletin, 85*, 1256-1274.

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1263-1274.

McArdle, J. J., & Anderson, E. (1990). Latent variable growth models for research on aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (3 ed., pp. 310-319). San Diego, CA: Academic Press.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2 ed.). London: Chapman and Hall.

McDonald-Miszczak, L., Hertzog, C., & Hultsch, D. F. (1995). Stability and accuracy of metamemory in adulthood and aging: A longitudinal analysis. *Psychology and Aging, 10*, 553-564.

McDonald-Miszczak, L., Hunter, M. A., & Hultsch, D. E. (1994). Adult age differences in predicting memory performance: The effects of normative information and task experience. *Canadian Journal of Experimental Psychology, 48*, 95-118.

McIntosh, A. R., Sekuler, A. B., Penpeci, C., Rajah, M. N., Grady, C. L., Sekuler, R., et al. (1999). Recruitment of unique neural systems to support visual memory in normal aging. *Current Biology, 9*, 1275.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*, 361-388.

Meeter, M., & Nelson, T. O. (2003). Multiple study trials and judgments of learning. *Acta Psychologica, 113*, 123-132.

Meiran, N., Israeli, A., Levi, H., & Grafi, R. (1994). Individual differences in self-reported cognitive failures: The attention hypothesis revisited. *Personality and Individual Differences, 17*, 727-739.

Merckelbach, H., Muris, P., Nijman, H., & de Jong, P. J. (1996). Self-reported cognitive failures and neurotic symptomatology. *Personality and Individual Differences, 20*, 715-724.

Merckelbach, H., Muris, P., & Rassin, E. (1999). Fantasy proneness and cognitive failures as correlates of dissociative experiences. *Personality and Individual Differences, 26*, 961-967.

Meredith, W. (1993). Meauserement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525-543.

Meredith, W., & Horn, J. L. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 203-240). Washington D.C.: American Psychological Association.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107-122.

Metcalfe, J. (1998). Cognitive optimism: Self-deception or memory-based processing heuristics? *Personality and Social Psychology Review, 2*, 100-110.

Metcalve, J., & Shimamura, A. (1994). *Metacognition: Knowing about knowing*. Cambridge, MA: Bradford Books.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*, 479-515.

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. New York: Springer.

Murphy, M. D., Sanders, R. E., Gabriesheski, A. S., & Schmitt, F. A. (1981). Metamemory in the aged. *Journal of Gerontology, 36*, 185-193.

Murphy, M. D., Sanders, R. E., Gabriesheski, A. S., & Schmitt, F. A. (1981). Metamemory in the aged. *Journal of Gerontology, 36*, 185-193.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2003). *Mx: Statistical modeling.* VCU Box 900126, Richmond, VA 23298: Department of Psychiatry. 6th Edition.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109-133.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51*, 102-116.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (jols) are extremely accurate at predicting subsequent recall: The "delayed-jol effect." *Psychological Science, 2*, 267-270.

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of swahili-english translation equivalents. *Memory, 2*, 325-335.

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 5*, 207-213.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1-25). Cambridge, MA: The MIT Press.

Nerb, J., Ritter, F. E., & Krems, J. (1999). Knowledge level learning and the power law: A soar model of skill acquisition in scheduling. *Kognitionswissenschaft, 8*, 20-29.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-51). Hillsdale, NJ: Erlbaum.

Newell, K. M., Liu, Y.-T., & Mayer-Kress, G. (2001). Time sclaes in motor learning and development. *Psychological Review, 108*, 57-82.

Niederehe, G., & Yoder, C. (1989). Metamemory perceptions in depressions of young and older adults. *Journal of Nervous and Mental Disease, 177*, 4-14.

Norman, D. A. (1981). Categorization of action slips. *Psychological Review, 88*, 1-15.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behaviour. In R. J. Davidson, Schwartz, G.E., Shapiro, D. (Ed.), *Consciousness and self regulation* (Vol. 4). New York: Plenum.

Nuttman-Schwartz, O. (2004). Like a high wave: Adjustment to retirement. *Gerontologist, 44*, 229-236.

O'Brien, R. M. (1985). The relationship between ordinal measures and their underlying values: Why all the disagreement? *Quality and Quantity, 19*, 265-277.

Oswald, W. D., & Fleischmann, U. M. (1999). *Nürnberger- alters-inventar* (4. ed.). Göttingen: Hogrefe.

Park, D. C., Lautenschlager, G., Hedden, T., Davison, N., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and Aging, 17*, 299-320.

Parr, W. V., & Siegert, R. (1993). Adults' conception of everyday memory failures in others: Factors that mediate the effects of target age. *Psychology and Aging, 8*, 599-605.

Paul, L. M. (1994). Making interpretable forgetting comparisons: Explicit versus hidden assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 992-999.

Pearman, A., & Storandt, M. (2004). Predictors of subjective memory in older adults. *Journal of Gerontology B: Psychological Sciences, 59*, 4-6.

Pedhazur, E. (1982). *Multiple regression in behavioural research: Explanation and prediction*. New York: Rinehart and Winston.

Perlmutter, M. (1978). What is memory aging the aging of? *Developmental Psychology, 14*, 330-345.

Pollina, L. K., Greene, A. L., Tunick, R. H., & Puckett, J. M. (1992). Dimensions of everyday memory in young adulthood. *British Journal of Psychology, 83*, 305-321.

Ponds, R. W. H. M., van Boxtel, M. P. J., & Jolles, J. (2000). Age-related changes in subjective cognitive functioning. *Educational Gerontology, 26*, 67-81.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing tom swift's electric factor analysis machine. *Understanding Statistics, 2*, 13-43.

Prull, M. W., Gabrieli, J. D. E., & Bunge, S. A. (1999). Age-related changes in memory: A cognitive neuroscience perspective. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of cognitive aging* (2 ed.). Mahwah, NJ: Erlbaum.

Quick, H. E., & Moen, P. (1998). Gender, employment, and retirement quality: A life course approach to the differential experiences of men and women. *Journal of Occupational Health Psychology*, 44-46.

Rabinowitz, J. C., Ackerman, B. P., Craik, F. I. M., & Hinchley, J. L. (1982). Aging and metamemory: The roles of relatedness and imagery. *Journal of Gerontology, 37*, 688-695.

Rast, P., & Zimprich, D. (submitted). Age Differences in the Underconfidence-With-Practice Effect. *Experimental Aging Research.*

Rast, P., Zimprich, D., Van Boxtel, M., & Jolles, J. (submitted). Factor structure and measurement invariance of the cognitive failures questionnaire across the adult life-span. *Psychological Assessment*.

Raz, N., Rodrigue, K. M., Head, D., Kennedy, K. M., & Acker, J. D. (2004). Differential aging of the medial temporal lobe: A study of a five-year change. *Neurology, 62*, 433-438.

Reason, J. (1988). Stress and cognitive failure. In S. Fisher & J. Reason (Eds.), *Handbook of life stress, cognition and health* (pp. 405-421). New York: Wiley.

Reason, J. (1990). *Human error*. England: Cambridge University Press.

Reason, J., & Lucas, D. (1984). Absent-mindedness in shops: Its incidence, correlates and consequences. *British Journal of Clinical Psychology, 23*, 121-131.

Rebok, G. W., & Balcerak, L. J. (1989). Memory self-efficacy and performance differences in young and old adults: The effect of mnemonic training. *Developmental Psychology, 25*, 714-721.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

Restle, F., & Greeno, J. G. (1970). *Introduction to mathematical psychology*. Reading, MA: Addison-Wesley.

Riby, L. M., Perfect, T. J., & Stollery, B. T. (2004). The effects of age and task domain on dual task performance: A meta-analysis. *European Journal of Cognitive Psychology, 16*, 863-891.

Richards, F. J. (1959). A flexible growth function for empirical use. *Journal of Experimental Botany, 10*, 290-300.

Richards, R. M., & Nelson, T. O. (2004). Effect of the difficulty of prior items on the magnitude of judgment of learning for subsequent items. *American Journal of Psychology, 117*, 81-91.

Ritter, F. E., & Schooler, L. J. (2001). The learning curve. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 8602-8605). Cambridge: Cambridge University Press.

Robinson-Whelen, S., & Kiecolt-Glaser, J. (1997). The importance of social versus temporal comparison appraisals among older adults. *Journal of Applied Social Psychology, 27*, 959-966.

Rönnlund, M., Nyberg, L., Bäckman, L., & Nilsson, L.-G. (2005). Stability, growth, and decline in adult life span development of declarative memory: Cross-sectional and longitudinal data from a population-based study. *Psychology and Aging, 20*, 3-18.

Royston, P., & Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Applied Statistics, 43*, 429-467.

Salthouse, T. A. (1991). *Theoretical perspectives on cognitive aging*. Hillsdale, NJ: Erlbaum.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review, 103*, 403-428.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507-514.

Schaie, K. W. (1994). The course of adult intellectual development. *American Psychologist, 49*, 304-313.

Schaie, K. W. (2005). *Developmental influences on adults intelligence: The seattle longitudinal study*. New York: Cambridge University Press.

Scheck, P., Meeter, M., & Nelson, T. O. (2004). Anchoring effects in the absolute accuracy of immediate versus delayed judgments of learning. *Journal of Memory and Language, 51*, 71-79.

Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect: Boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General, 134*, 124-128.

Schneider, W. (1985). Developmental trends in the metamemory-memory behavior relationship: An integrated review. In D. L. Forrest-Pressley, G. E. MacKinnon & T. G. Waller (Eds.), *Cognition, metacognition, and human performance* (Vol. 1, pp. 57-109). New York: Academic Press.

Schneider, W., & Pressley, M. C. (1989). *Memory development between 2 and 20*. New York: Springer-Verlag.

Schneider, W., Visé, M., Lockl, K., & Nelson, T. O. (2000). Developmental trends in children's memory monitoring. Evidence from a judgment-of-learning (jol) task. *Cognitive Development, 15*.

Schoenberg, M. R., Dawson, K. A., Duff, K., Patton, D., Scott, J. G., & Adams, R. L. (2006). Test performance and classification statistics for the rey auditory verbal learning test in selected clinical samples. *Archives in Clinical Neuropsychology, 21*, 693-703.

Serra, M. J., & Dunlosky, J. (2005). Does retrieval fluency contribute to the underconfidence-with-practice effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1258-1266.

Shaw, R. J., & Craik, F. I. M. (1989). Age differences in predictions and performance on a cued recall task. *Psychology and Aging, 4*, 131-135.

Sidman, M. (1952). A note on functional relations obtained from group data. *Psychological Bulletin, 49*, 263-269.

Silvia, P. J., & Gendolla, G. H. E. (2001). On introspection and self-perception: Does self-focused attention enable accurate self-knowledge? *Review of General Psychology, 5*, 241-269.

Smith, G. (1998). Learning statistics by doing statistics. *Journal of Statistics Education [Online]*, 6.

Smith, G., Sala, S. D., Logie, R. H., & Maylor, E. A. (2000). Prospective and retrospective memory in normal ageing and dementia: A questionnaire study. *Memory, 8*, 311-321.

Stern, Y. (2002). What is cognitive reserve? Theory and research application of the reserve concept. *Journal of the International Neuropsychological Society, 8*, 448-460.

Suls, J., & Mullen, B. (1983-1984). Social and temporal bases of self-evaluation in the elderly: Theory and evidence. *International Journal of Aging and Human Development, 18*, 111-120.

Sunderland, A., Harris, J. E., & Baddeley, A. D. (1983). Do laboratory tests predict everyday memory? A neuropsychological study. *Journal of Verbal Learning and Verbal Behavior, 22*, 341-357.

Swaminathan, H., & Algina, J. (1978). Scale freeness in factor analysis. *Psychometrika, 43*, 581-583.

Thurstone, L. L. (1919). The learning curve equation. *Psychological Monographs, 26*, 1-51.

Touron, D. R., & Hertzog, C. (2004). Distinguishing age differences in knowledge, strategy use, and confidence during strategic skill acquisition. *Psychology and Aging, 19*, 452-466.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*, 371-386.

Tulving, E. (1964). Intratrial and intertrial retention: Notes towards a theory of free recall verbal learning. *Psychological Review, 71*, 219-237.

Tulving, E. (2001). Episodic memory and common sense: How far apart? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 356*, 1505-1515.

Tulving, E., & Madigan, S. A. (1970). Memory and verbal learning. *Annual Review of Psychology, 21*, 437-484.

Vandenberg, R. J. (2002). Towards a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*, 139-159.

Verhaeghen, P., Marcoen, A., & Goossens, L. (1993). Facts and fiction about memory aging: A quantitative integration of research findings. *Journal of Gerontology, 48*, 157–171.

Vom Hofe, A., Mainemarre, G., & Vannier, L. (1998). Sensitivity to everyday failures and cognitive inhibition: Are they related? *European Review of Applied Psychology, 48*, 49-55.

Wagle, A. C., Berrios, G. E., & Ho, L. (1999). The cognitive failures questionnaire in psychiatry. *Comprehensive Psychiatry, 40*, 478-484.

Wallace, J. C. (2004). Confirmatory factor analysis of the cognitive failures questionnaire: Evidence for dimensionality and construct validity. *Personality and Individual Differences, 37*, 307-324.

Wallace, J. C., Kass, S. J., & Stanny, C. J. (2002). The cognitive failures questionnaire revisited: Dimensions and correlates. *The Journal of General Psychology, 129*, 238-256.

Wallace, J. C., & Vodanovich, S. J. (2003). Can accidents and industrial mishaps be predicted? Further investigation into the relationship between cognitive failure and reports of accidents. *Journal of Business and Psychology, 17*, 503-514.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review, 20*, 158.

Wellman, H. M. (1983). Metamemory revisited. In M. T. H. Chi (Ed.), *Trends in memory development research* (pp. 31-51). Basel, Switzerland: Karger.

Wellman, H. M. (1985). The origins of metacognition. In D. L. Forrest-Pressley, G. E. MacKinnon & T. G. Waller (Eds.), *Metacognition, cognition, and human performance* (pp. 1-31). New York: Academic Press.

Wicklund, R. A., & Eckert, M. (1992). *The self-knower: A hero under control*. New York: Plenum Press.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance abuse domain. In K. J. Bryant & M. Windle (Eds.), *The science of prevention: Methodological advance from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Williams, J. M., Little, M. M., Scates, S., & Blockman, N. (1987). Memory complaints and abilities among depressed older adults. *Journal of Consulting and Clinical Psychology, 55*, 595-598.

Willis, S. L., & Schaie, K. W. (2005). Cognitive trajectories in midlife and cognitive functioning in old age. In S. L. Willis & M. Martin (Eds.), *Middle adulthood: A lifespan perspective* (pp. 243-276). Thousand Oaks, CA: Sage.

Willis, S. L., Tennstedt, S. L., Marsiske, M., Ball, K., Elias, J., Mann Koepke, K., et al. (2006). Long-term effects of cognitive training on everyday functional outcomes in older adults. *Journal of the American Medical Association, 296*, 2805-2814.

Wisher, R. A., Sabol, M. A., & Kern, R. P. (1995). Modeling acquisition of an advanced skill: The case of morse code copying. *Instructional Science, 23*, 381-403.

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30*, 165-200.

Zacks, R. T., Hasher, L., & Li, K. Z. H. (2000). Human memory. In T. Salthouse & F. I. M. Craik (Eds.), *The handbook of aging and cognition* (2 ed., pp. 293-357). Mahwah, NJ: Lawrence Erlbaum Associates.

Zelinski, E. M., Burnight, K. P., & Lane, C. J. (2001). The relationship between subjective and objective memory in the oldest old: Comparison of findings from a representative and a convenience sample. *Journal of Aging and Health, 13*, 248-266.

Zimprich, D. (2002). Cross-sectionally and longitudinally balanced effects of processing speed on intellectual abilities. *Experimental Aging Research*, 231-251.

Zimprich, D., Allemand, M., & Hornung, R. (2006). Measurement invariance of the abridged sense of coherence scale in adolescence. *European Journal of the Psychological Assessment, 22*, 280-287.

Zimprich, D., Allemand, M., & Huber, S. (2007). *Measurement invariance in ordered categorical variables: The case of big-five personality markers in adulthood.* Manuscript submitted for publication.

Zimprich, D., Hofer, S. M., & Aartsen, M. J. (2004). Short-term versus long-term longitudinal changes in processing speed. *Gerontology, 50*, 17-21.

Zimprich, D., & Kliegel, M. (in press). An age-comparative analysis of predictors of cognitive complaints in middle and old adulthood. *Journal of Adult Development*.

Zimprich, D., Kliegel, M., Dellenbach, M., Rast, P., Zeintl, M., & Martin, M. (in revision). Cognitive ability structure in old age: First results from the zurich longitudinal study on cognitive aging. *Swiss Journal of Psychology*.

Zimprich, D., & Martin, M. (2002). Can longitudinal changes in processing speed explain longitudinal changes in fluid intelligence? *Psychology and Aging, 17*, 690-695.

Zimprich, D., Martin, M., & Kliegel, M. (2003). Subjective cognitive complaints, memory performance, and depressive affect in old age: A change-oriented approach. *International Journal of Aging and Human Development, 57*, 339-366.

Zimprich, D., Perren, S., & Hornung, R. (2005). A two-level confirmatory factor analysis of a modified rosenberg self-esteem scale. *Educational and Psychological Measurement, 65*, 465-481.

Zimprich, D., Rast, P., & Martin, M. (in press). Individual differences in verbal learning in old age. In S. M. Hofer & D. F. Alwin (Eds.), *The handbook of cognitive aging: Interdisciplinary perspectives*. Thousand Oaks: Sage Publications.

Appendix

Modelling Description

In what follows, we will shortly describe the factor model of ordered-categorical variables and how, based on this model, measurement invariance of ordered-categorical variables across groups may be examined.

Let Y_{ijk} denote the score on the *j*th ordered-categorical measure for the *i*th person in the *k*th group. In the factor model for ordered-categorical data, the observed scores Y_{ijk} are assumed to be determined by unobserved scores on latent response variates Y^*_{ijk}. Other than the observed measures Y_{ijk}, these latent response variates are continuous in scale (Millsap & Yun-Tein, 2004). The observed measures can be viewed as discretized versions of the latent response variates, given that scores on the observed measures are determined through

$$Y_{ijk} = m \quad \text{if} \quad v_{jkm} \leq Y^*_{ijk} < v_{ij(m+1)}, \tag{1}$$

where m = 0, 1, ..., c categories are confined by c + 1 $\{v_{jk0}, v_{jk1}, ..., v_{jk(c+1)}\}$ latent threshold parameters for the *j*th variable as measured on persons from the *k*th group. The items of the Cognitive Failures Questionnaire (CFQ), after having collapsed the category "often" and "very often," discretize the latent responses into c = 4 four categories. The two extreme thresholds are pre-defined: $v_{jk0} = -\infty$ and $v_{jk(c+1)} = +\infty$. The remaining c threshold parameters may vary across variables and across groups (Millsap & Yun-Tein, 2004).

The probabilities associated with observed values for Y_{ijk} are determined by the probability distribution of latent response variate Y^*_{ijk}. Let $\mathbf{Y}^T_{ik} = \{Y_{i1k}, Y_{i2k}, ..., Y_{ipk}\}$ be the $1 \times p$ vector of observed scores on the p variables for the *i*th person in the *k*th group, with $\mathbf{Y}^{*T}_{ik}$ the analogous vector of scores on the latent response variates. It is typically assumed that

$$\mathbf{Y}_{ik} \sim MVN(\boldsymbol{\mu}^*_k, \boldsymbol{\Sigma}^*_k), \tag{2}$$

where $\boldsymbol{\mu}^*_k$ is a $p \times 1$ vector of means on the latent response variates, and $\boldsymbol{\Sigma}^*_k$ is a $p \times p$ covariance matrix for the latent response variates, each subscripted to permit differences in

these parameters between groups. Note that the multivariate normal distribution is assumed for the latent response variates and not for the observed ordered categorical variables. Given the latent response variates Y_{ijk}^{*}, the factor model is specified for these variates as

$$Y_{ijk}^{*} = \tau_{jk} + \boldsymbol{\lambda}_{jk}^{T}\boldsymbol{\eta}_{ik} + e_{ijk}, \qquad (3)$$

where τ_{jk} is a latent intercept parameter, $\boldsymbol{\lambda}_{jk}$ is an $r \times 1$ vector of factor loadings for the *j*th variate on *r* factors, $\boldsymbol{\eta}_{ik}$ is the $r \times 1$ vector of factor scores for the *i*th person in the *k*th group, and e_{ijk} is the *j*th unique factor score for that person. Letting $\mathbf{e}_{ik}^{T} = \{e_{i1k}, e_{i2k}, ..., e_{ipk}\}$ be the $1 \times p$ vector of unique factor scores, we assume that

$$\boldsymbol{\eta}_{ik} \sim MVN(\boldsymbol{\alpha}_{k}, \boldsymbol{\Psi}_{k}), \quad \mathbf{e}_{ik} \sim MVN(\mathbf{0}, \boldsymbol{\Theta}_{k}), \qquad (4)$$

with $\boldsymbol{\alpha}_{k}$ an $r \times 1$ vector of factor means, $\boldsymbol{\Psi}_{k}$ an $r \times r$ factor covariance matrix, and $\boldsymbol{\Theta}_{k}$ a $p \times p$ diagonal covariance matrix for the unique factors. We also assume that $Cov(\boldsymbol{\eta}_{ik}, \mathbf{e}_{ik}) = \mathbf{0}$ for all *i*, *k* (Millsap & Yun-Tein, 2004). These assumptions lead to the structure

$$E(\mathbf{Y}_{ik}^{*}) = \boldsymbol{\mu}_{k}^{*} = \boldsymbol{\tau}_{k} + \boldsymbol{\Lambda}_{k}\boldsymbol{\alpha}_{k}, \quad Cov(\mathbf{Y}_{ik}^{*}) = \boldsymbol{\Sigma}_{k}^{*} = \boldsymbol{\Lambda}_{k}\boldsymbol{\Psi}_{k}\boldsymbol{\Lambda}_{k}^{T} + \boldsymbol{\Theta}_{k}, \qquad (5)$$

where $\boldsymbol{\tau}_{k}^{T} = \{\tau_{1k}, \tau_{2k}, ..., \tau_{pk}\}$ and $\boldsymbol{\Lambda}_{k}$ is a $p \times r$ factor pattern matrix whose *j*th row is $\boldsymbol{\lambda}_{jk}^{*T}$. All factor model parameters are subscripted to permit differences. In sum, the analysis of ordered categorical variables results in finding an adequate structure, which is assumed to stem from a multivariate normal distribution, underlying the ordered categorical variables. The main problem is one of model identification, because the scores on measured variables are only indirectly determined, and typically $\boldsymbol{\mu}_{k}^{*}$ and $\boldsymbol{\Sigma}_{k}^{*}$ are not themselves identified.

Factorial invariance and model identification

As Millsap and Yun-Tein (2004) stated, the relevant parameters for factorial invariance in multiple groups with ordered-categorical measures, are the thresholds $\{v_{jk0}, v_{jk1}, ..., v_{jk(c+1)}\}$ and the factor model parameters $(\boldsymbol{\tau}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\Theta}_k)$.

To identify factor models based on ordered-categorical measures in multiple groups, an additional set of constraints on the model parameters is needed, compared to the ordinary factor analysis model for continuous measures. Millsap and Yun-Tein (2004) proposed to approach the identification problem by two steps: First, constraints are needed on the latent responses and on the variate parameters $(\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*)$ within each group. If they are identified and estimable, constraints are imposed on the factor model parameters $\{\boldsymbol{\tau}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\Theta}_k, \boldsymbol{\Psi}_k, \boldsymbol{\alpha}_k\}$. Furthermore two cases are to be distinguished when models are to be identified. Models can be of a congeneric or of a noncongeneric structure. In the first case, suppose that in the factor model in equation 5, each row of $\boldsymbol{\Lambda}_k$ has only one nonzero element. This would indicate that each latent response variate loads only on one factor. In the second, more complex case, variables load on multiple factors, for example, the three factor solution we presented here allows Item 2 of the CFQ to load on all factors. Millsap and Yun-Tein (2004) proposed five constraints which are sufficient to achieve identification of a noncongeneric model in multiple groups when c > 1 and r >1 (for the congeneric case see Millsap & Yun-Tein, 2004). To identify all threshold parameters in one group it is sufficient to fix $\boldsymbol{\mu}_k^* = \mathbf{0}$ and $Diag(\boldsymbol{\Sigma}_k^*) = \mathbf{I}$. Next, $\boldsymbol{\alpha}_k = \mathbf{0}$ is to be fixed in this same group. Then, in all groups the latent intercept parameters are fixed to $\boldsymbol{\tau}_k = \mathbf{0}$. Additionally, constraints are imposed on $\boldsymbol{\Lambda}_k$ to be rendered rotationally unique within each group. In the fourth step two values of *m* are to be chosen, and for each of the chosen values $v_{jkm} = v_{jm}$ for all *k*, with *j* = 1, …, *p*. These constraints force two thresholds per measurement variable to be invariant. These constraints are sufficient to render the baseline model of configural invariance identifiable. Because the subsequent models of measurement invariance impose more constraints, they are also identified.

Short Curriculum Vitae

Education

2004-2007	University of Zurich, Department of Psychology, Gerontopsychology, doctoral candidate
1998-2004	University of Berne, Department of Psychology, licentiate (lic. phil.) in General- and Neuropsychology, Minor: Sociology

Employment History

Since 2004	Research assistant at the Department of Psychology, Gerontopsychology, University of Zurich